REPORT RESUMES

ED 016 263 24 CG 001 356

THE USE OF THE COMPUTER TO GENERATE STATISTICAL TABLES FOR THE STUDY OF PERSONALITY TRAITS, A MONTE CARLO AND A LOGICAL ANALYSIS OF MULTITRAIT-MULTIMETHOD STATISTICS.

BY- JACOBSON, MILTON D.
VIRGINIA UNIV., CHARLOTTESVILLE, SCH. OF EDUC.

REPORT NUMBER BR-5-8410 PUB DATE JUN 67
EDRS PRICE MF-$0.50 HC-$4.20 103P.

DESCRIPTORS- *DATA ANALYSIS, *COMPUTERS, *MATHEMATICAL LOGIC, *MATHEMATICAL APPLICATIONS, *PERSONALITY ASSESSMENT, TABLES (DATA), TEST VALIDITY, CORRELATION MATRICES,

THIS WAS A TWO PART INVESTIGATION. THE FIRST PART WAS A MONTE CARLO (STATISTICAL) ANALYSIS, AND THE SECOND WAS A LOGICAL ANALYSIS OF MULTITRAIT-MULTIMETHOD VALIDITY. PART ONE SUCCESSFULLY GENERATED, FOR SMALL SAMPLE SIZES, EMPIRICAL DISTRIBUTIONS OF STANLEY'S F STATISTIC FOR TESTING TRAIT VALIDITY IN MULTITRAIT-MULTIMETHOD MATRICES. ZYZANSKI'S CORRECTION OF STANLEY'S STATISTIC WAS FOUND INAPPROPRIATE FOR SMALL SAMPLE SIZES. PART TWO EMPLOYED THE METHOD OF LOGICAL ANALYSIS TO DETERMINE THE SOUNDNESS OF THE FOUR CRITERIA PROPOSED BY CAMPBELL AND FISKE FOR DETERMINING TRAIT VALIDITY BY MULTITRAIT-MULTIMETHOD MATRICES. IT WAS CONCLUDED THAT ONLY CRITERION ONE, CONVERGENT VALIDITY, COULD BE CONSIDERED A "THEOREM" OF TESTING THEORY. THIS ANALYSIS QUESTIONED WHETHER SPECIFIC TESTS CAN BE VALIDATED OR INVALIDATED WHEN THE CRITERIA OFFERED TO DO THIS ARE NOT THEMSELVES "VALID" OR LOGICALLY NECESSARY. (PS)

ED016263

THE USE OF THE COMPUTER TO GENERATE STATISTICAL TABLES FOR THE STUDY OF PERSONALITY TRAITS:

A MONTE CARLO AND LOGICAL ANALYSIS

Cooperative Research Project No. 5-8410

MILTON D. JACOBSON

THE UNIVERSITY OF VIRGINIA
BUREAU OF EDUCATIONAL RESEARCH

5-8410
24

THE USE OF THE COMPUTER TO GENERATE STATISTICAL TABLES FOR THE STUDY OF PERSONALITY TRAITS:

(A Monte Carlo and a Logical Analysis of Multitrait-Multimethod Statistics)

Contract No. 5-8410

Principal Investigator:
Milton D. Jacobson

June, 1967

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Bureau of Educational Research
University of Virginia

Charlottesville, Virginia

CG 001 356

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ACKNOWLEDGEMENTS

This research was made possible by the stimulation and guidance of Leroy Wolins (Professor of Statistics and Psychology at Iowa State University) with whom the project director served as an NSF postdoctoral fellow,

preliminary discussions on methodology with Raymond O. Collier, Jr. (Professor of Educational Psychology at the University of Minnesota) and Desmond S. Cartwright (Professor of Psychology at the University of Colorado),

the utility of several computer subroutines which were an integral part of the computer program system and which were developed and made available by Robert A. Bottenberg and Joe H. Ward, Jr. (Personnel Research Activity, Lackland Field, San Antonio, Texas),

the analytical skills of John Young (NDEA Fellow in Symbolic Logic and Doctoral Student at the University of Virginia) who played a major role in the development of the logical analysis in interactions with the investigator and James T. Cargile (Assistant Professor of Philosophy at the University of Virginia),

the energy and skill of Robert McCarthy, my undergraduate assistant, who has a way with computers,

the support of the National Science Foundation, the University of Virginia, and the Cooperative Research Program of the Office of Education, U.S. Department of Health, Education, and Welfare which funded this project,

the facilities of the Bureau of Educational Research directed by Mary Ann MacDougall and the expertise of its secretaries Mrs. Judith Hockey and Mrs. Lynn Smith who prepared this report.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF TABLES & GRAPHS v

INTRODUCTION 1

Part I Monte Carlo Analysis of the Statistical (Probabilistic) Problem for Small Sample Sizes 2

Objectives of Part I 3

Part II Logical Analysis 4

Objectives of Part II 4

Related Literature 5

Figure 1. Results from factorial studies 15

Figure 2. Results from split plot studies 16

METHOD 18

Part I Monte Carlo Analysis of the Statistical (Probabilistic) Problem for Small Sample Sizes 19

Part II Logical Analysis 21

RESULTS 23

Part I Monte Carlo Analysis 24

Stanley's F_{PT} Statistic 25

Zyzanski's F_{PT} Statistic 40

Part II Logical Analysis 41

I. Introduction 41

II. A "Good" Test 45

III. The Meaning of "Valid" 48

IV. Campbell & Fiske: Towards a Definition of a Good Test 51

(A) Multitrait-Multimethod Approach to Reliability. . 56
(B) Validity and the Matrix 59
(1) Convergent Validity: Criterion I 59
(2) Discriminant Validity: Criterion II-IV . . . 65
Criterion II 66
Criterion III 67
Criterion IV 69
V. Conclusion 70
DISCUSSION 78
Part I Monte Carlo Analysis 79
Part II Logical Analysis 79
CONCLUSIONS AND IMPLICATIONS 81
Part I Monte Carlo Analysis 82
Implications 84
SUMMARY 85
Part I Monte Carlo Analysis 86
Stanley's F_{PT} Statistic 87
Results 87
I. Purpose of our Investigation 88
II. Method of Inquiry 89
III. Conclusions 89
REFERENCES (and Notes) 91
APPENDIX I 93
APPENDIX II 94

LIST OF TABLES AND GRAPHS

TABLES FOR EVALUATING THE ROBUSTNESS OF F_{PT} STATISTICS FOR NON-NULL CONTRIBUTIONS OF METHOD (b_j) AND METHOD-TRAIT (W_{jk}) BIAS 27

Table **Page**

1 For 2 Methods and 2 Traits 28
2 For 2 Methods and 3 Traits 29
3 For 2 Methods and 4 Traits 30
4 For 2 Methods and 5 Traits 31
5 For 3 Methods and 3 Traits 32
6 For 3 Methods and 4 Traits 33
7 For 3 Methods and 5 Traits 34
8 For 4 Methods and 4 Traits 35
9 For 4 Methods and 5 Traits 36
10 Summary of the Weightings of Method (b_j) and Method-trait (W_{jk}) Which Minimize Chi Square Values Best 39
11 Best Weightings of Method (b_j) and Method-trait (W_{jk}) Bias 83

Graph

1 Convergence of X^2 for Varying Weightings of Method (b_j) bias 37
2 Convergence of X^2 for Varying Weightings of Method-trait (W_{jk}) bias 38

INTRODUCTION

Campbell and Fiske advocated the use of multitrait-multimethod intercorrelation matrices and developed four criteria which provide for both convergent and discriminant validation of psychological traits. Their criteria were found in the more than fifty year history of test and measurement literature. Their paper tried to bridge the gap between atheoretical practices and theoretical formulations in measurement.

Briefly, the validational process emphasized

1) convergent validity and its distinction from reliability,

2) discriminant validity in which methods of measurement can be invalidated by high correlations with other methods from which they should differ,

3) trait-method units in which each trait is considered in combination with methods not restricted to the measurement of that trait, and

4) the necessity of measuring more than one trait (multitrait) by more then one method (multimethod).[1]

Campbell and Fiske recognized logical difficulties and statistical (probabilistic) difficulties in multitrait-multimethod validation.[2] It was the purpose of this research to investigate these two difficulties, and these have been treated separately.

Part I Monte Carlo Analysis of the Statistical (Probabilistic) Problem for Small Sample Sizes

This research investigated the appropriateness of using the statistics developed for these intercorrelation matrices to validate data obtained from small sample sizes.

Although statistical theory dictates the distribution function of certain statistics, given a set of assumptions, such theory will rarely reveal the distribution of the statistics when one or more of the assumptions are violated. Moreover, it is often impossible to obtain the distribution by analytical methods. Under these conditions it is useful to determine the distribution of the statistic by means of Monte Carlo procedures. This methodology typically employs an electronic computer to generate a large number of computed values of a statistic. The computer is programmed to sample from populations whose parameters are known, and the distribution of a statistic is studied as a function of the parameters of a given population. A purpose of this research was to use Monte Carlo procedures to obtain, for small sample sizes, empirical distributions of certain F-statistics which may be calculated for the problem defined by Campbell and Fiske (1) in their article entitled "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix."

Originally, analyses of multitrait-multimethod correlation matrices were made without objective summary statistical procedures. Derivations of such statistics were made by Stanley (8) and Zyzanski (10) using three-way factorial designs where the three factors were persons, methods and traits.

In person-method-trait studies trait validity is usually estimated by the variance component attributable to person-by-trait interaction effect, and invalidity may arise from four possible sources of method bias which are usually estimated by the variance components attributable to: method (halo) effect, person-by-method interaction effect, method-by-trait interaction effect (error), and person-by-method-by-trait interaction effect.

The robustness of the F statistic used to determine trait validity (person-trait interaction effect) was evaluated for various combinations of non-null contributions of the four sources of method bias for small sample sizes.

Stanley's statistic was developed to provide a probabalistic interpretation of Campbell and Fiske's multitrait-multimethod intercorrelation matrix. Zyzanski's statistic is similar to Stanley's, but provides for and permits the analysis of data which are normally encountered in day to day educational measurement practice. Such measurements do not have comparable reliabilities. Thus, it was expected that the two statistics when generated empirically would disagree. This disagreement was noted and compared with Campbell and Fiske's criteria to determine the usefulness of each statistic under the practical conditions being considered.

Objectives of Part I

1. To generate for small sample sizes, empirical distributions of the F statistics (Stanley's and Zyzanski's) for testing trait validity in a multitrait-multimethod matrix.
2. To determine if these statistics remain invariant for various combinations of non-null contributions of the sources of method and error bias.
3. To compare Stanley's statistic with Zyzanski's and with the criteria of Campbell and Fiske.
4. If necessary, to present the prescribed conditions which permit the use of these statistics.

Part II Logical Analysis

The difficulty with the statistical treatment of multitrait-multimethod matrices which Zyzanski (10) identified could have been due to Stanley's (8) simplification of assumptions, or it could have been due to logical difficulties in Campbell and Fiske's formulation of their criteria. Thus a logical analysis was made of Campbell and Fiske's four criteria which were: convergent validity, discriminant validity, trait-method unit and the multitrait-multimethod requirement. Campbell and Fiske recognized difficulties with the last two criteria when they stated that, "...our insistence on more than one method for measuring each concept departs from Bridgeman's early position that 'if we have more than one set of operations, we have more than one concept, and strictly there should be a separate name to correspond to each different set of operations'."(1)

This analysis of these criteria are logically necessary or merely contingently necessary. It also attempted to clarify the interrelationship between the criteria. Arguments were made in ordinary language and in symbolic language.

Objectives of Part II

1. Determine if Campbell and Fiske's criteria are contingently or logically necessary, and
2. to clarify the interrelationships between Campbell and Fiske's criteria.

3. Related Literature.

A behavioral scientist who uses correlation coefficients to explore the relationship between two variables is working with a two-stage sampling scheme (6,7). He samples people from a population to which he will generalize, and within each person he samples from two populations of responses, one for each variable. A few examples of populations of responses (traits), defined by behavioral scientists are general intelligence, verbal fluency, quantitative reasoning, introversion-extroversion, sociability and dominance. Each person's test-score is a composite score obtained by summing the score for each response.

In all psychological measuring devices, certain features are introduced specifically to represent the trait that it is intended to measure. There are other features characteristic of the method being employed, and these features could also be present in efforts to measure other quite different traits. The test, or rating scale, or other device, almost always elicits systematic variance in response due to both groups of features. To the extent that irrelavant method variance or systematic person-method interaction bias contributes to the scores obtained, these scores are invalid.

This source of invalidity has been identified in the literature since 1920 and has been described as halo effects in studies of ratings, as apparatus factors in animal studies, and as response sets or test form factors in paper and pencil tests (1). Halo effects bear the responsibility for "causing" such nonsensical relationships as the correlation (.63) between the quality of voice and teacher's intelligence. Apparatus factors pre-empt psychological factors and are exemplified by the correlation (.87) between measurements of hunger and thirst in an activity wheel (different constructs measured by same method) being of the same magnitude as their test-retest reliability (.83 and .92 respectively). Test-form factors represent variance due to item format (multiple choice, true-false, etc.), IBM answer sheets, variability in the subjects' conscientiousness, motivation, or test-taking sophistication and are often confused and confounded with a "general test factor" (1).

Campbell and Fiske (1) advocate a validational process utilizing a matrix of intercorrelations among trait measurements which represent at least two traits, each measured by at least two methods. Measures of the same trait should correlate higher with each other than they do with measures of different traits by different methods. Theoretically, these monotrait-heteromethod validity values should be higher than correlations among different traits measured by different methods. If the monotrait-heteromethod values are higher than the heteromethod-heterotrait values, one may attribute unique variance to at least one of the traits. Thus a trait with unique variance has potential for predicting criteria for which it is rationally relevant.

In applying the multitrait-multimethod validational procedure to experimental data taken from the literature Campbell and Fiske found that the preceding desirable conditions, as a set, are rarely met. They summarize their findings as follows:(1)

> "Multitrait-multimethod matrices are rare in the test and measurement literature. Most frequent are two types of fragment: two methods and one trait (single isolated values from the validity diagonal, perhaps accompanied by a reliability or two), and heterotrait-monomethod triangles. Either type of fragment is apt to disguise the inadequacy of our present measurement efforts, particularly in failing to call attention to the preponderant strength of methods variance."
>
> "The illustrations of multitrait-multimethod matrices presented so far give a rather sorry picture of the validity of the measures of individual differences involved. The typical case shows an excessive amount of methods variance, which usually exceeds the amount of trait variance. This picture is certainly not as a result of a deliberate effort to select shockingly bad examples; these are ones we have encountered without attempting an exhaustive coverage of the literature. The several unpublished studies of which we are aware show the same picture. If they seem more disappointing than the general run of validity data reported in the journals, this impression may very well be because the portrait of validity provided by isolated values plucked from the validity diagonal is deceptive, and uninterpretable in isolation from the total matrix."

Campbell and Fiske have made a strong case for validation by means of the multitrait-multimethod correlation matrix. Their arguments include illustrating its use in research studies, its theoretical and empirical agreement with previous formulations, such as, construct validity and convergent operationalism, and its improvement over other methods in directing an experimenter towards gains over preceding stages of his work in measurement by specifically indicating which methods should be discarded or which concepts are poorly measured because of excessive or confounded method variance. This indicated action for the experimenter can be determined by a careful examination of an appropriate multitrait-multimethod matrix.

Campbell and Fiske also consider the problem of developing summary statistical procedures for use when determining "valid variance" by means of the multitrait-multimethod matrix because applications of their criteria to data contain exceptions, and the exact demarcation point which distinguishes trivial exceptions from significant exceptions is blurred. Thus, the development of objective probabalistic statistical procedures necessary for an improved analysis of the multitrait-multimethod matrix they left to future investigators (1).

Derivations of such objective summary statistics have been made using three way factorial designs where the factors are persons, methods, and traits (8),(10). Before discussing their work a preliminary treatment of the mean squares attributable to the effects in an ordinary three way factorial analysis will be made in order to establish notation which can be used throughout this report.... and to establish their relationship to the measurement of validity and invalidity.

The variance is assigned to the three main effects of persons, methods and traits, to the three first order interaction effects of person by method, person by trait and method by trait and to the second order interaction effect of person by method by trait. The mean squares of these seven effects will be respectively denoted by MS_P, MS_M, MS_T, MS_{PM}, MS_{PT}, MS_{MT} and MS_{PMT}.

Invalidity due to method bias is usually determined from the three mean squares involving method MS_M, MS_{PM} and MS_{MT}. These may reflect, respectively, differences among some methods in general level of rating, bias of some methods toward certain individuals, and bias of some methods toward certain traits (3), (9).

Willingham and Jones (9) also related validity to the component MS_{PT}, which reflects differential meaning of the various traits. Valid variance in person-method-trait studies is usually determined from this MS_{PT} component. Validity might also be determined from the MS_P and MS_T components, but these are less frequently used.

The MS_{PM} component is independent of the MS_{PT} component

(1), (4). Thus, in any one study one may find any degree of relative method (halo) effect and any degree of trait independence, and in multitrait-multimethod matrices these two mean squares constitute separate criteria for the adequacy of ratings.

Stanley (8) developed statistics to test for invalidity and validity as measured by these two mean squares. The three mean squares needed for these test statistics were derived from the three way factorial design and are expressed in terms of covariances in equations (1), (2) and (3).

$$(1)\ MS_{PM} = \frac{\sum_{k=k'}^{M} \sum_{j,j'}^{T} \frac{C_{jk,\,j'k'}}{T} - \sum_{j,j'}^{T} \sum_{k,k'}^{M} \frac{C_{jk,\,j'k'}}{T\,M}}{(P-1)\ (M-1)}$$

$$(2)\ MS_{PT} = \frac{\sum_{j=j'}^{T} \sum_{k,k'}^{M} \frac{C_{jk,\,j'k'}}{M} - \sum_{j,j'}^{T} \sum_{k,k'}^{M} \frac{C_{jk,\,j'k'}}{T\,M}}{(T-1)\ P-1)}$$

$$(3)\ MS_{PMT} =$$

$$\frac{\sum_{j=j'}^{T} \sum_{k=k'}^{M} C_{jk,\,j'k'} - \sum_{k=k'}^{M} \sum_{j,j'}^{T} \frac{C_{jk,j'k'}}{T} - \sum_{j=j'}^{T} \sum_{k,k'}^{M} \frac{C_{jk,j'k'}}{M} + \sum_{j,j'}^{T} \sum_{k,k'}^{M} \frac{C_{jk,j'k'}}{TM}}{(P-1)(T-1)(M-1)}$$

Equations (4) and (5) give the test statistics.

(4) (invalidity)

$$F_{PM} = \frac{MS_{PM}}{MS_{PMT}} = \frac{\text{equation (1)}}{\text{equation (3)}}$$

(5) (validity)

$$F_{PT} = \frac{MS_{PT}}{MS_{PMT}} = \frac{\text{equation (2)}}{\text{equation (3)}}$$

In summary of the previous works, Campbell and Fiske had developed a validational procedure using a matrix of intercorrelations obtained from measuring people on at least two traits by at least two methods. Analysis of such a matrix of correlations provides the experimenter with a measurement of the validity of the traits he is measuring and of the degree of method bias. In addition this analysis indicates to the experimenter the direction he should take to improve trait validity and to reduce method bias. Campbell and Fiske's work lacked objective summary statistics. Stanley derived these statistics (F) from a three-way factorial design where the factors are persons, methods, and traits. Stanley's F statistics to determine validity and method bias were obtained by an analysis using covariances. There is a gap between the work of Campbell and Fiske and that of Stanley because covariances and correlations are not identical. If one is to use correlations (and this seems desirable) one must assume comparable reliability (non-heterogeneity of correlation) (1) among all tests in order to assign the method variance to the monomethod and heteromethod blocks in the correlation matrix. This assumption is often violated by real test and measurement data. Zyzanski contributed a theoretical correction for data for which this assumption is not fulfilled so that it would be possible to make both a probabalistic analysis like Stanley's and an inspection analysis like Campbell and Fiske's on all measurement data.

Zyzanski's work is now presented in two parts, theoretical and empirical. In the theoretical part the rationale and mathematical development of his statistics will be sketched. In the empirical section the evidence based on the application of his analysis to experimental data taken from twenty-five studies in the literature is presented to show its substantial agreement with the conventional analysis.

Theoretical. Zyzanski derived the following equation from correlation and reliability theory.

$$(6) \quad \frac{C_{jk,j'k'}}{S_{jk}, S_{j'k'}(P-1)} = \overset{c}{r}_{jk,j'k'} = r_{jk,j'k'} \sqrt{\frac{S_{jk}^2 + S_{ejk}^2}{S_{jk}^2}} \sqrt{\frac{S_{j'k'}^2 + S_{ej'k'}^2}{S_{j'k'}^2}}$$

j - method, k - trait
jk - trait-method combination

$\dot{r}_{jk,j'k'}$	is an estimate of the correlation coefficient for measures on method-trait combination jk with method-trait combination j'k' overall people. This estimate is made by means of the split half reliability and estimates the correlation which would occur if twice as many items were included in the test.
$r_{jk,j'k'}$	is the actual correlation coefficient which is calculated.
S^2_{jk}	is the estimate of the variance of the method-trait interaction effect.
(P-1)	is the degrees of freedom for persons.
$C_{jk,j'k'}$	$= \sum_i X_{ijk} X_{ij'k'} - \frac{\sum_i X_{ijk} \sum_i X_{ij'k'}}{P}$ and if divided by P-1 represents the covariance corresponding to $r_{jk,j'k'}$ above.

P-1 is the degrees of freedom for persons.

Inspection of equations (1), (2) and (3) reveals that the exact F statistics require only one term, $C_{jk,j'k'}$ (summations of this term are made in four different ways however). Equation (6) relates this term, $C_{jk,j'k'}$, to two other terms, $\dot{r}_{jk,j'k'}$ and $r_{jk,j'k'}$.

Conceivably, the analysis could be made using any of these three terms. Previous investigators Lord (5) and Cochran (2) suggest doing the analysis on $\dot{r}_{jk,j'k'}$ values and referring the results to a chi-square table. In a three-way factorial this would require the assumption of constant error variance ($E(S^2_{MPT}) = 0$ or σ^2_{MPT}). Zyzanski did not make this assumption because it restricts the analysis to large groups of people.

Zyzanski wished to use the correlation term $r_{jk,j'k'}$ and he derived a procedure which permits the

(7) $E(S^2_{MPT}) = \sigma_e^2$ or $\sigma_{MPT}^2 + \sigma_e^2$

Zyzanski proposes to deal with the part of equation (6) expressed in equation (8)

$$(8) \qquad \frac{C_{jk,j'k'}}{S_{jk}S_{j'k'}\,(P=1)} = r_{jk,j'k'}\sqrt{\frac{S_{jk}^2 + S_{ejk}^2}{S_{jk}^2}}\sqrt{\frac{S_{j'k'}^2 + S_{ej'k'}^2}{S_{j'k'}^2}}$$

Since S_{jk}^2 and S_{ejk}^2 are confounded in equation (8) both must be constant to permit an analysis. Zyzanski successfully treated this confounded error term by assuming one of its parts constant and mathematically adjusting the other part so it appears constant. After a consideration of the type of psychological data described by Campbell and Fiske, he concluded that it was better to assume S_{jk}^2 constant than to assume S_{ejk}^2 constant. He assumed S_{jk}^2 constant and equal to unity because analysis of variance is not subject to a scale transformation. Thus by this assumption, equation (8) is simplified to

$$(9) \qquad \frac{C_{jk,j'k'}}{(P-1)} = r_{jk,j'k'}\left[\sqrt{1 + S_{ejk}^2}\sqrt{1 + S_{ej'k'}^2}\right]$$

Next he proposed a mathematical adjustment of equation (9) which makes S_{ejk}^2 and $S_{ej'k'}^2$ constant. he used the Spearman-Brown prophecy formula given in equation (10) to adjust the $r_{jk,j'k'}$ values for unequal reliabilities as determined from the split-half reliability procedure.

$$(10) \qquad r_{jk,j'k'} = \dot{r}_{jk,j'k'}\sqrt{\dot{r}_{jk,jk}}\sqrt{\dot{r}_{j'k',j'k'}}$$

$$\dot{r}_{jk,j'k'} = \frac{S_{jk}^2}{\sqrt{S_{jk}^2 + S_{ejk}^2}}$$

$\dot{r}_{jj}$	is an estimate of the reliability of test $_j$.
$\dot{r}_{kk}$	is an estimate of the reliability of test $_k$.
S_j	is the standard deviation of people on jth measure.
S_k	is the standard deviation of people on kth measure.
S_{ej}^2	is the error variance associated with test $_j$.
S_{ek}^2	is the error variance associated with test $_k$.
$\dot{r}_{jk}$	is an estimate of correlation if twice as many items had been used.

A specific example of this procedure is shown below.

Example Given: $r_{jk} = .50$, $\dot{r}_{jj} = .70$, $\dot{r}_{kk} = .80$

Estimate what r_{jk} would be if $\dot{r}_{jj} = .35$, $\dot{r}_{kk} = .40$

Answers: Using 5

$$\dot{r}_{jk} = \frac{.50}{\sqrt{.70}\ \sqrt{.80}}$$

$$\text{so } r_{jk} \text{ (estimated)} = r_{jk}\sqrt{.40}\ \sqrt{.35} = .25$$

Thus equation (10) permits one to estimate the correlation between two variables that would result for arbitrary values of the reliabilities. In equation (10) Zyzanski equated the reliabilities, $\dot{r}_{jk,jk}$ and $\dot{r}_{j'k',j'k'}$, and set them equal to r_a. Then he substituted the right side of equation (10) for the term, $r_{jk,j'k'}$ in equation (9) giving

$$(11)\quad \frac{C_{jk,j'k'}}{(P-1)} = \dot{r}_{jk,j'k'}\underset{A}{\left[\sqrt{\dot{r}_{jk,jk}\ \dot{r}_{j'k',j'k'}}\right]}\underset{B}{\left[\sqrt{1 + S_{ejk}^2}\right]}$$

Terms A and B of equation (11) must cancel in order that the equality from equation (6) can hold. After simplification this results in

$$(12) \qquad r_a = \frac{1}{r_{jk,jk}} = \frac{1}{1+S_{ejk}^{2}}$$

Zyzanski proposes to use the term r_a to obtain by means of equation (13) adjusted correlation coefficients $r_{jk,j'k'}$ which are estimates of what the correlation between measurements jk and j'k' would be if both were subject to the average error.

$$(13) \qquad r'_{jk,j'k'} = r_{jk,j'k'} \; r_a = \left[\frac{C'_{jk,j'k'}}{(P-1)}\right]$$

Thus Zyzanski developed a procedure which permitted the adjustment of a correlation matrix to account for heterogeneity of reliability (or unequal errors of measurement) and allowed one to use either the adjusted correlations or the adjusted covariances to get an approximate F statistic. The development of this statistic required the assumption of constant trait-method interaction variance ($E(S_{jk}^{2})$ = constant) and the statistic is approximate because of this assumption.

If correlations were used, the test statistics are given by equations (14) and (15).

$$(14) \qquad F_{(TP)} = \frac{(\bar{r}_{wt} - \bar{r}_0)\ (M-1)}{1 - \bar{r}_{wm} - \bar{r}_{wt} + \bar{r}_0}$$

$$(15) \qquad F_{(MP)} = \frac{(\bar{r}_{wm} - \bar{r}_0)\ (T-1)}{1 - \bar{r}_{wm} - \bar{r}_{wt} + \bar{r}_0}$$

$\bar{r}_{wm}$ is the average correlation within methods

$\bar{r}_{wt}$ is the average correlation within traits

$\bar{r}_{o}$ is the average overall correlation.

$F'_{(MP)}$ person method interaction F

$F'_{(TP)}$ is trait person interaction F

Empirical. Zyzanski collected data from sixteen studies reported in the literature from 1959 to 1961 where three-way factorial analyses had been carried out and analyzed it by his procedure. If the data had more than one observation per subclass the correlations were corrected and the analyses were carried out both with and without the correction for unequal error variance. The results of these analyses were compared with those from the conventional analysis to determine their agreement.

Zyzanski portrays the results of the agreement between F' and the F required for significance at the five per cent level in Figure 1 by plotting the ratio of the two F's. Those values which fall in the lower left quadrant (below 1.00) indicate agreement between insignificant values for both F's. Those that fall in the upper right quadrant (above 1.00) indicate agreement between significant values for both F's.

Inspection of Figure 1 reveals that the agreement between the two F's is substantial. There are, however, 5 cases where the discrepancy was large enough to cause only one of the two F values to be significant. Investigation of these revealed that the data possessed certain deviations, such as the small degrees of freedom which explain the contrary results. Four of the five cases were from replicated studies. In these studies the analyses were done with and without the correction and the correction brought the approximate F in closer agreement with the theoretical F.

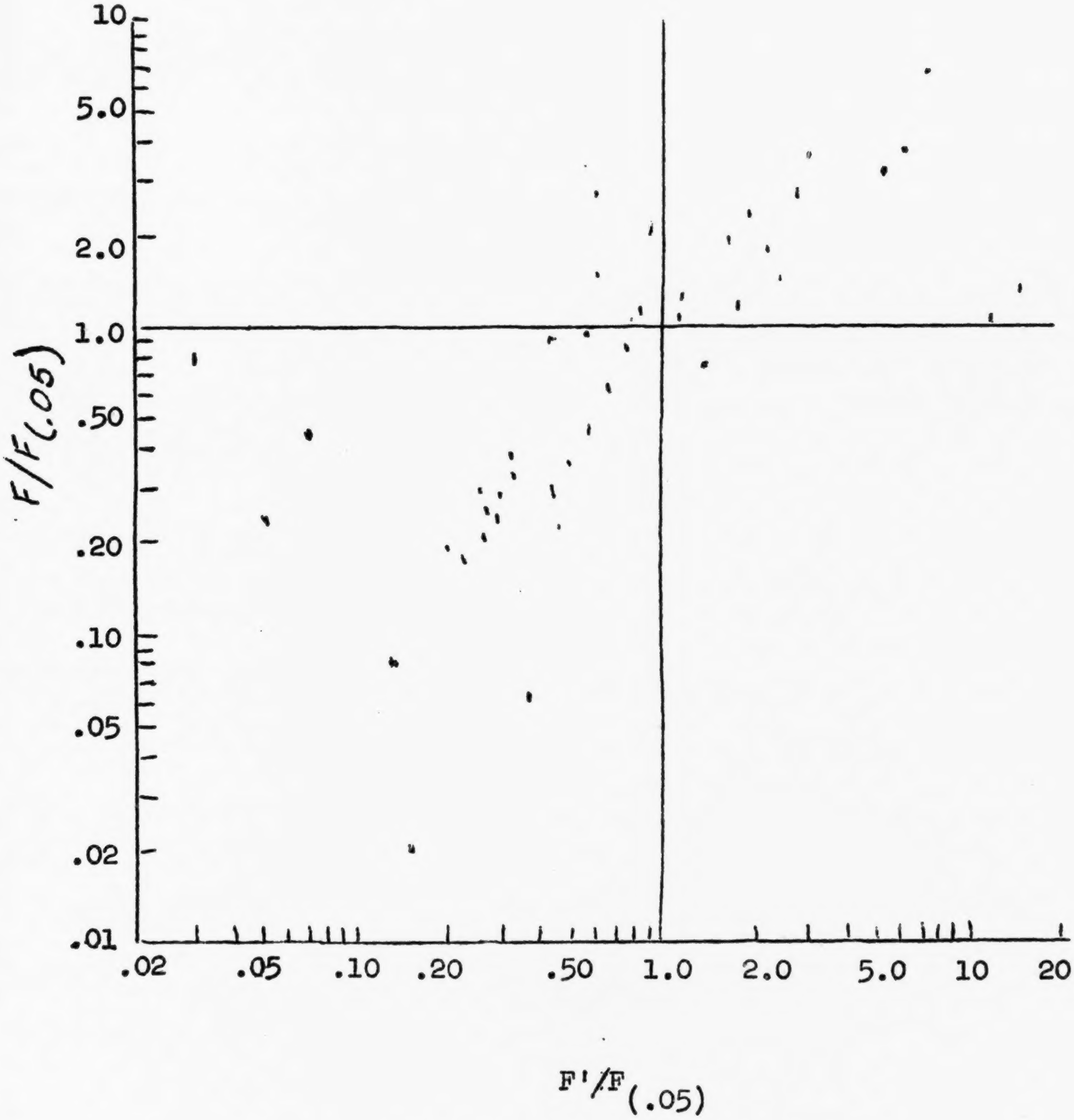


Figure 1. Results from factorial studies

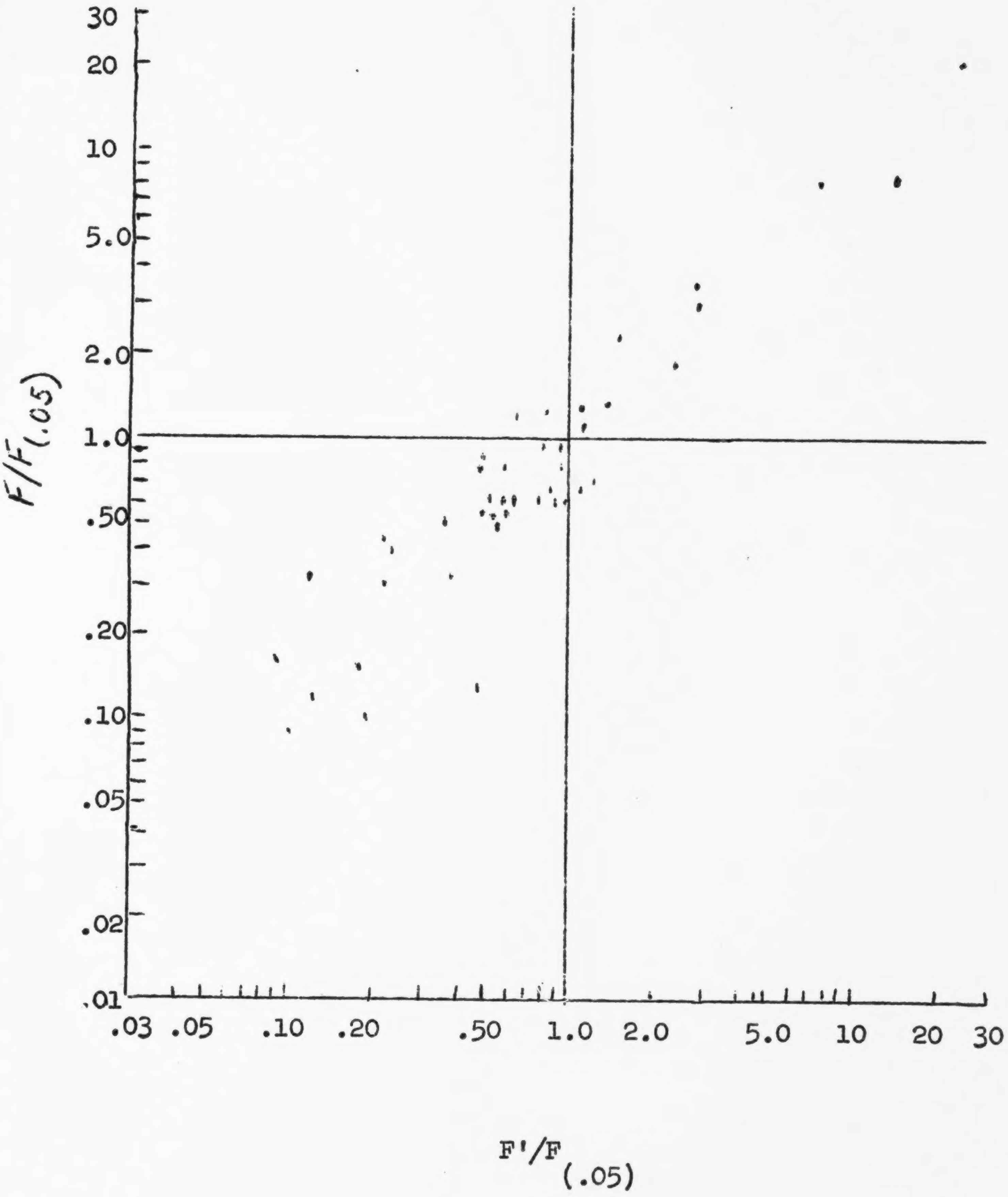


Figure 2. Results from split plot studies

Zyzanski analytically extended this approximate procedure to split plot studies in which the analysis had been made on factors without comparable scales. Nine studies in the literature were analyzed and the results reported in Figure 2. Again the agreement between the approximate F and the theoretical F is considerable. There are five discrepancies which are, however, of less magnitude than were those of the factorial studies. Once again the magnitude of the discrepancies seemed to be related to the number of degrees of freedom.

A summary of Zyzanski's work shows that the approximate F's which he developed allowed one to adjust an intercorrelation matrix in order to permit analyses by means of correlation coefficient of data which do not fulfill the assumption of comparable reliabilities. Zyzanski's work supplemented and intimately related the previous work of Stanley with that of Campbell and Fiske. The usefulness of the approximate analysis of Zyzanski derives from the following reasons.

1. It permits one to inspect and analyze a correlation matrix by Campbell and Fiske's method and by a probabalistic one similar to Stanley's even though the assumption of comparable reliabilities is violated.

2. It provides statistics which give a general and substantial agreement with those from a theoretical or exact F analysis.

3. It produced a promising extension into measurement areas where comparable scales were not available.

METHOD

Part I Monte Carlo Analysis of the Statistical (Probabilistic) Problem for Small Sample Sizes.

Although statistical theory dictates the distribution function of certain statistics, given a set of assumptions, such theory will rarely reveal the distribution of the statistic when one or more of the assumptions are violated. Moreover, it is often impossible to obtain the distribution by analytical methods. Under these conditions it is useful to determine the distribution of the statistic by means of Monte Carlo procedures typically employing an electronic computer to generate a large number of computed values of a statistic. The computer is programmed to sample from populations whose parameters are known, and the distribution of a statistic is studied as a function of the parameters of a given population. This research used Monte Carlo procedures to obtain, for small sample sizes, and a small number of non-null conditions empirical distributions of the Stanley (8) and Zyzanski (10) $F_{(PT)}$ statistic. This statistic is the most important statistic for the determination of validity in a person-method-trait study.

Monte Carlo procedures are feasible only when the investigator is interested in the distribution of the statistic under null conditions or a small number of non-null conditions. In this research three null conditions were considered. These were due to the variance effects attributable to person-method, person-method-trait, and to persons. The mean square of the first of these effects, MS_{PM}, was shown to be independent of the MS_{PT} whose distribution we determined ((1), (4)) and it was considered as a random effect operating under null conditions. Ther person-method-trait effect was considered as one of two terms confounded in the error variance, but the correlation (or covariance) values on which this analysis was based were adjusted to remove contributions of this effect, and it also was considered as a random effect operating under null conditions. The third effect due to persons, MS_p, was considered to be operating under null conditions because the random selection of persons is assumed in every experiment of the multitrait-multimethod variety from which one might wish to generalize (1) to other populations.

The variance effects attributable to the method-trait interaction, MS_{MT}, and to the method main effect, MS_M, were considered as operating under non-null conditions. As the method-trait interaction variance, MS_{MT}, was assumed by others (8) (10), to be constant and equal to unity, it was determined under what conditions, if any, violations of this assumption affect the distribution of the approximate $F_{(PT)}$ statistic. Of all the variance components the method variance MS_M is most likely to make a non-null contribution and a review of the measurement literature from 1920 to the present revealed the obvious truth of this statement and underlined the importance of determining the effects, if any, that a non-null method variance would induce in the distribution of the approximate $F_{(PT)}$ statistic.

In order to calculate the statistic for testing the significance of the person-trait interaction effect (8),(10), a matrix of intercorrelations of P persons scores on M methods and T traits was generated. The ith person's score on the jth method and the kth trait was created by summing three random variables each of which were a sample drawn by Monte Carlo procedure from a normal univariate distribution with zero mean and unit variance. These three variables represented random effects (null conditions) for person, i, person-method interaction, ij, and person-method-trait interaction (error), ijk. Each person-method interaction variable was multiplied by a wighting factor which was determined for each method and was constant over the T traits and P persons. Each person-method-trait interaction variable was multiplied by a weighting factor which was determined for each of the MT method-trait interactions and was constant over the P persons.

The mathematical model for obtaining the PMT scores is given in equation 16.

$$(16) \quad X_{ijk} = P_i + b_j \, m_{ij} + w_{jk} \, e_{ijk}$$

The two weighting factors (b_j and w_{jk}) were related by restricting the average correlation over persons ($X_{ijk} \; X_{ij'k'}$) to the three categories, low ($\overline{r} = .3$), medium ($\overline{r} = .7$), and high ($\overline{r} = .9$). Theoretically the weighting factors and the correlation are related by equation 17.

$$(17) \quad \rho(X_{ijk}X_{ij'k'}) = \frac{i \quad 1}{\sqrt{1 + b_j^2 + w_{jk}^2} \; \sqrt{1 + b_{j'}^2 + w_{j'k'}^2}}$$

The term r_a which was used for Zyzanski's statistic to adjust the sample estimates, $r_{ijk,ij'k'}$, of $\rho_{ijk,ij'k'}$, the population coefficient, was determined by means of equation (12) and for this model was

$$(18) \quad r_a = \frac{1}{\overline{1/r}_{ijk,ij'k'}} = \frac{1}{1 + \overline{(1+b_j^2 + w_{jk}^2)}}$$

Once the PMT scores had been obtained these were correlated over persons to give an MT by MT intercorrelation matrix. Stanley and Zyzanski's F statistics for testing person-trait-interaction were calculated for this matrix using both adjusted (Zyzanski) and unadjusted (Stanley) correlation coefficients. The entire procedure for obtaining this matrix and statistic was repeated 1000 times. This gave an empirical distribution, with 1000 points for each statistic. P (sample size) was varied from 5 to 30 and for these sample sizes M (number of methods) was varied from 2 to 5, T (number of traits) was varied from 2 to 5 and the average correlation was restricted to the three values .3, .7, and .9. 150 empirical distributions were generated.

The effect on the empirical distributions of the variations in method and error (method-trait) variances described were determined by using the chi square test for goodness of fit. The observed frequencies of each empirical distribution with 1000 points were compared with the expected frequencies of the F distribution in the categories in the cumulative distribution function limited by 0.0 to .90, .90 to .95, .95 to .98, .98 to .99 and .99 to infinity.

If the chi square value was too large the weights b_j and W_{jk} were adjusted and empirical F_{PT} statistics were again generated until the Chi square values converged to a minimum. This feeding back and updating of the Monte Carlo procedure resulted in the prescription of limits within which the affected sources of variation could be analyzed by means of the F statistics considered. Statistical Tables for these prescribed limits are presented.

Part II Logical Analysis

The logical analysis was limited in scope and, of course, in method. What we attempted to accomplish was a critical examination of the four criteria presented by Campbell and Fiske (1) to determine the grounds which justify our acceptance, and/or use, of these criteria. It involves taking certain ideas we have about 1) what a test is and 2) what a good test should do, and relating these common sense concepts of "test," "validity" and "reliability" to the concepts of "test," "validity" and "reliability" as used by Campbell and Fiske and other people working in the field of psychological testing.

We compared the two sets of concepts and tried to determine whether set I was compatible with Set II, or whether the relationship between I and II was one of entailment, contradiction, etc. In short, the venture was strictly analytic and logical in the technical senses of these two words. We made no attempt to discover what *does* occur in the realm of the empirical nor even to predict what should occur. We simply worked on the basis of what is logically possible (i.e., not self-contradictory).

The purpose of this procedure was to examine the foundations, the very roots of testing theory. Just as we ask of a test, "Is it trustworthy, and if so, why so?", so too we must ask of our criterion, "Is it trust-worthy, and if so, why so?" We cannot make sound judgments, if our norms for gauging valid tests are wrong or misleading. Consequently, we must ask of theorists like Campbell and Fiske, "Are your criteria sound, and if so, why so?"

The method, as we mentioned above, was not experimental or inductive, but deductive and *a priori*. We tried, on the basis of an ordinary language analysis, as well as an analysis transformed into symbolic logic, to see whether the criteria of Campbell and Fiske were *entailed* by our common sense demands on testing. That is, *could one deduce these criteria as logically necessary conclusions*, from certain notions of "test," etc. The method we employed required neither praise nor condemnation of any results achieved. It is entirely expository and clarificatory, not evaluative. To say that the criteria could (not) be deduced is only to say that they are (not) theorems, as it were, derivable from prior axioms. This tells us only *what kind* of statements are made, *not* what the statements are worth.

RESULTS

This research investigated the appropriateness of using multitrait-multimethod intercorrelation matrices and Campbell and Fiske's criteria (1) as a validational process. This was a two part investigation, statistical and logical, and these were treated separately, and the results are reported separately.

PART I MONTE CARLO ANALYSIS

The Monte Carlo analyses investigated the multitrait-multimethod intercorrelation matrices to validate data obtained from small sample sizes. These statistics were developed by Stanley (8) and Zyzanski (10) using three-way factorial designs where the three factors were persons, methods, and traits.

The Objectives of this part of the study were:

1. To generate for small sample sizes, empirical distributions of the F statistics (Stanley's and Zyzanski (10) for testing trait validity in a multitrait-multimethod matrix.
2. To determine if these statistics remain invariant for various combinations of non-null contributions of the sources of method and error bias.
3. To compare Stanley's statistic with Zyzanski's and with the criteria of Campbell and Fiske.
4. If necessary, to present the prescribed conditions which permit the use of these statistics.

Objectives 1 and 2 were achieved by the following prodedures. The mathematical model for obtaining the Person-Method-Trait scores is given in equation 16.

$$(16) \quad X_{ijk} = P_i + b_j m_{ij} + w_{jk} e_{ijk}$$

In equation 16 the terms P_i, m_i, and e_{ijk} were random normal numbers generated on the computer,* and represent null conditions as described in the Method chapter. The non-null conditions were represented by the terms b_j and W_{jk} which were treated as two weighting factors. P_i represented each persons variability. The other terms, b_j, m_{ij}, w_{jk} and e_{ijk} represented the four possible sources of method bias which are estimated by variance components attributable to: method (halo) effect (b_j), person-by-method interaction effect (m_{ij}), method-trait interaction effect (W_{jk}), and person-by-method-by-trait interaction effect (e_{ijk}).

---

* The selection of the random normal number generator is described in appendix 1.

The two weighting factors (b_j and w_{jk}) were related by restricting the average correlation over persons ($\rho x_{ijk}, x_{ij'k'}$) to three categories, low ($\bar{r}$ = .3), medium ($\bar{r}$ = .7), and high ($\bar{r}$ = .9_. Theoretically the weighting factors and the correlation are related by equation 17.

$$(17)\ \bar{\rho}\,(x_{ijk} x_{ij'k'}) = \sum_i \frac{1}{\sqrt{1 + b^2_j + w^2_{jk}}\ \sqrt{1 + b^2_{j'} + w^2_{j'k'}}}$$

The weighting factors were restricted to specific degrees of inequality and to specific proportions of total variance which they contributed and were determined for the three values of ρ by means of equation 17.

Once the Person-Method-Trait, PMT, scores were obtained these were correlated over persons to give an MT by MT (M is the number of Methods and T the number of Traits) intercorrelation matrix. Both Stanley's and Zyzanski's F statistics for testing person-trait-interaction were calculated for this matrix using both adjusted (Zyzanski's) and unadjusted (Stanley's) correlation coefficients. The entire procedure for obtaining this matrix and statistic were repeated 1000 times. This gave an empirical distribution with 1000 points for each statistic.

Stanley's F_{PT} Statistic

Approximately 150 such empirical distributions were generated. Each empirical distribution was compared with its theoretical F distribution with the chi-squared goodness of fit test. The results of these comparisons are given in Tables 1 through 9. Each table reports data for one particular combination of M and T (eg. Table 1, M=2, T=2, Table 2, M=2, T=3). In each table the sample size, P, is listed. The theoretical and empirical correlation values are also listed except for cases where empirical values were not calculated. The weighting factors due to method (b_j) and method-trait (W_{jk}) are also listed. The sixth column in each table lists the chi-squared (x^2) value for those cases in which it was calculated. Chi-squared values and empirical correlations were not calculated for empirical distributions which contained more than 100 negative F values since negative F values are not theoretically possible.

The success with which the first objective of this research was achieved can be determined by comparing the empirical and theoretical correlation values in these tables. Close agreement between these values indicates successful completion of this objective. Each empirical correlation is an average of the 1000 correlation values each of which came from averaging the MT by MT correlations in each matrix.

The degree of invariance of the statistics (Objectives two) can be determined by inspecting column 6 in these tables. The smaller the chi-squared value reported in column 6 the more invariant are the statistics for the non-null conditions described in columns 4 and 5.

The data in Tables 1 through 9 demonstrates that Stanley's F_{PT} statistic is not invariant or robust under non-null conditions of method (b_j) and method-trait (W_{jk}) bias.

The chi-square (X^2) values for a good fit of the empirical F to the theoretical F should be less than 9.49 (5 per cent significance level). The chi-square values in tables 1-9 vary from 9.96 to more than 100,000 as the contributions of method (b_j) and method-trait (W_{jk}) bias are varied.

By modifying the weights b_j and W_{jk} it was possible to obtain minimum chi square values. This is shown in Graph 1 where several cases taken from Tables 1-9 have been plotted (chi square value versus weight b_j). Since specifications of b_j also specifies W_{jk} it is redundant to show a plot of chi square and W_{jk}, but this is shown in graph 2 for clarity only.

Those weightings of method (b_j) and method-trait (W_{jk}) which give minimal chi square values are presented in Table 10. In all but a few cases it is clearly shown what the best combinations of weightings are.

TABLES FOR EVALUATING THE ROBUSTNESS OF F_{PT} STATISTICS

FOR NON-NULL CONTRIBUTIONS OF METHOD (b_j) AND METHOD-TRAIT (W_{jk}) BIAS.

TABLES 1 - 9

TABLE 1

For 2 Methods and 2 Traits. M=2, T=2.

P	$\bar{\rho}$ Emp.	$\bar{\rho}$ Theor.	b	W	X^2
5	*	0.7	2/3	1/3	*
5	*	0.7	1/2	1/2	*
5	*	0.7	8/100	92/100	*
5	*	0.9	8/100	92/100	*
5	*	0.7	1/12	11/12	*
5	*	0.7	1/10	9/10	*
5	*	0.7	1/6	5/6	*
5	*	0.7	1/5	4/5	*
5	*	0.7	3/10	7/10	*
5	*	0.7	1/3	2/3	*
5	*	0.7	2/5	3/5	*
5	*	0.7	5/12	7/12	*
5	*	0.9	3/10	7/10	*
15	0.66	0.7	5/100	95/100	107
15	0.66	0.7	7/100	93/100	90.4
15	0.67	0.7	8/100	92/100	81.3
15	0.67	0.7	9/100	91/100	106
15	0.68	0.7	1/10	9/10	108
15	0.68	0.7	12/100	88/100	102
15	0.69	0.7	14/100	86/100	126
15	0.699	0.7	16/100	84/100	160
15	0.71	0.7	18/100	82/100	141
15	0.71	0.7	2/10	8/10	170
15	0.76	0.7	1/3	2/3	843
15	0.78	0.7	2/5	3/5	2611
15	0.87	0.9	8/100	92/100	99
25	0.68	0.7	8/100	92/100	32.7
25	0.87	0.9	8/100	92/100	42.3

* > 100 neg F's

TABLE 2

For 2 Methods and 3 Traits. M=2, T=3.

P	$\bar{\rho}$ Emp.	$\bar{\rho}$ Theor.	b	W	X^2
5	*	0.7	1/2	1/2	*
5	0.66	0.7	1/10	9/10	67.8
5	0.68	0.7	13/100	87/100	64.1
5	0.689	0.7	16/100	84/100	57.8
5	0.704	0.7	19/100	81/100	54.7
5	0.708	0.7	2/10	8/10	52.6
5	0.715	0.7	22/100	78/100	56.2
5	0.723	0.7	24/100	76/100	52.4
5	0.73	0.7	1/4	3/4	48.5
5	0.73	0.7	26/100	74/100	49.1
5	0.74	0.7	28/100	72/100	52.8
5	0.75	0.7	3/10	7/10	57.6
5	*	0.7	4/10	6/10	*
10	0.467	0.3	1/4	3/4	74.5
10	0.709	0.7	1/4	3/4	75.1
10	0.905	0.9	1/4	3/4	67.1
20	0.71	0.7	1/4	3/4	88.2

* $>$100 neg. F's

TABLE 3

For 2 Methods and 4 Traits. M=2, T=4.

P	$\bar{\rho}$ Emp.	$\bar{\rho}$ Theor.	b	W	X^2
5	0.59	0.7	1/10	9/10	93.7
5	0.635	0.7	2/10	8/10	81
5	0.678	0.7	3/10	7/10	62.3
5	0.709	0.7	38/100	62/100	28.5
5	0.71	0.7	39/100	61/100	30.6
5	0.716	0.7	4/10	6/10	49.9
5	0.72	0.7	41/100	59/100	97.6
5	0.72	0.7	42/100	58/100	149

TABLE 4

For 2 Methods and 5 Traits. M=2, T=5.

P	$\bar{\rho}_{\text{Emp.}}$	$\bar{\rho}_{\text{Theor.}}$	b	W	X^2
5	0.577	0.7	1/10	9/10	109
5	0.62	0.7	2/10	8/10	93
5	0.667	0.7	3/10	7/10	71
5	0.6997	0.7	38/100	62/100	36
5	0.707	0.7	4/10	6/10	57
5	0.72	0.7	42/100	58/100	134
5	0.72	0.7	44/100	56/100	277
5	0.744	0.7	1/2	1/2	985

TABLE 5

For 3 Methods and 3 Traits. M=3, T=3.

P	$\bar{\rho}$ Emp.	$\bar{\rho}$ Theor.	b	W	X^2
5	*	0.7	1/10	9/10	*
5	*	0.7	2/10	8/10	*
5	*	0.7	3/10	7/10	*
5	*	0.7	4/10	6/10	*
10	0.602	0.7	1/10	9/10	106
10	0.62	0.7	15/100	85/100	109
10	0.64	0.7	2/10	8/10	110
10	0.65	0.7	22/100	78/100	110
10	0.67	0.7	27/100	73/100	112
10	0.68	0.7	3/10	7/10	113
10	0.71	0.7	4/10	6/10	114
10	0.71	0.7	42/100	58/100	111

* $>$ 100 neg. F's

TABLE 6

For 3 Methods and 4 Traits. M=3, T=4.

P	$\bar{\rho}_{\text{Emp.}}$	$\bar{\rho}_{\text{Theor.}}$	b	W	X^2
5	0.33	0.3	1/10	9/10	24.3
5	0.679	0.7	1/3	2/3	10.4
5	0.88	0.9	1/3	2/3	31.9
15	0.705	0.7	1/3	2/3	12.98
15	0.895	0.9	1/3	2/3	9.96
25	*	0.7	2/3	1/3	*
25	0.75	0.7	1/2	1/2	17753
25	0.71	0.7	1/3	2/3	17.05
30	*	0.9	2/3	1/3	*
30	0.91	0.9	1/2	1/2	21533
30	0.901	0.9	1/3	2/3	18.69

* $>$100 neg. F's

TABLE 7

For 3 Methods and 5 Traits. M=3, T=5.

P	$\bar{\rho}_{\text{Emp.}}$	$\bar{\rho}_{\text{Theor.}}$	b	W	X^2
5	0.56	0.7	1/10	9/10	89
5	0.602	0.7	2/10	8/10	90
5	0.64	0.7	3/10	7/10	87
5	0.669	0.7	38/100	62/100	69.6
5	0.675	0.7	4/10	6/10	59.5
5	0.68	0.7	42/100	58/100	54.2
5	0.685	0.7	44/100	56/100	53.9
5	0.6889	0.7	46/100	54/100	54.2
5	0.6966	0.7	1/2	1/2	72.9

TABLE 8

For 4 Methods and 4 Traits. M=4, T=4.

P	$\bar{\rho}_{Emp.}$	$\bar{\rho}_{Theor.}$	b	W	X^2
20	0.486	0.3	1/2	1/2	10227
5	*	0.7	7/3	1/3	*
5	0.727	0.7	1/2	1/2	14090
5	0.56	0.7	1/14	13/14	159
5	0.57	0.7	1/12	11/12	162
5	0.57	0.7	1/11	10/11	169
5	0.58	0.7	1/10	9/10	167
5	0.587	0.7	1/8	7/8	186
5	0.595	0.7	1/7	6/7	210
5	0.675	0.7	1/3	2/3	1328
5	0.68	0.7	4/10	6/10	3018
5	*	0.9	2/3	1/3	*
5	0.925	0.9	1/2	1/2	14067
5	0.889	0.9	1/3	2/3	1499
10	0.328	0.3	1/10	9/10	162
10	0.727	0.7	1/2	1/2	41840
10	0.689	0.7	1/3	2/3	2797
10	0.578	0.7	1/16	15/16	202
10	0.904	0.9	1/2	1/2	40538
10	0.888	0.9	1/3	2/3	3021
20	0.358	0.3	1/6	5/6	522
20	0.576	0.7	1/26	25/26	325
20	0.6999	0.7	1/3	2/3	6311
20	0.91	0.9	1/2	1/2	101526
20	0.86	0.9	1/6	5/6	749

* $>$100 neg. F's

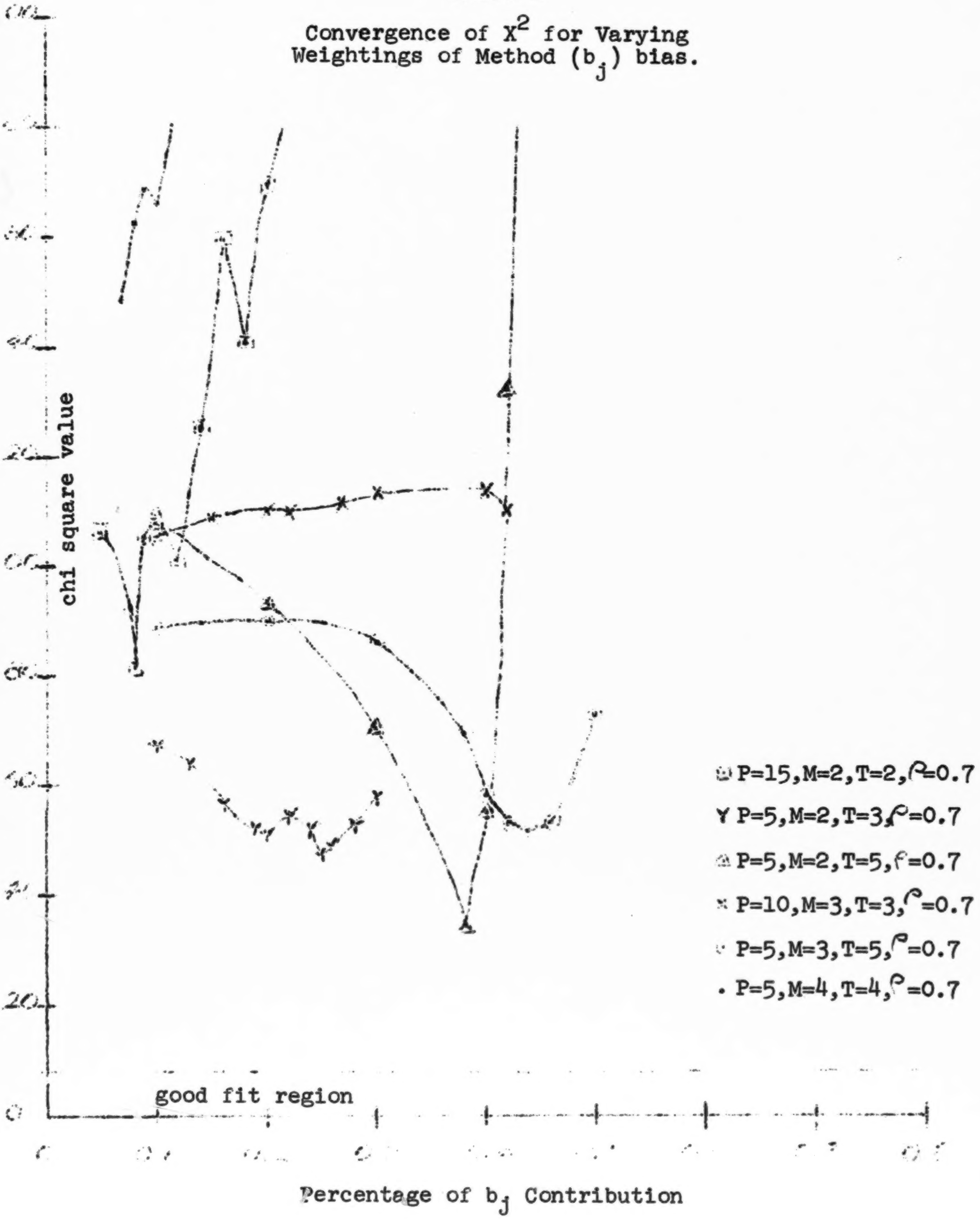
TABLE 9

For 4 Methods and 5 Traits. M=4, T=5.

P	$\bar{\ }$ Emp.	$\bar{\ }$ Theor.	b	W	X^2
5	0.548	0.7	8/100	92/100	129
5	0.56	0.7	1/10	9/10	140
5	0.575	0.7	14/100	86/100	177
5	0.603	0.7	2/10	8/10	251
5	0.644	0.7	3/10	7/10	775
5	0.676	0.7	4/10	6/10	4036
5	0.6986	0.7	1/2	1/2	16557

GRAPH 1

Convergence of X^2 for Varying Weightings of Method (b_j) bias.



GRAPH 2

Convergence of X^2 for Varying Weightings of Method-trait (W_{jk}) Bias.

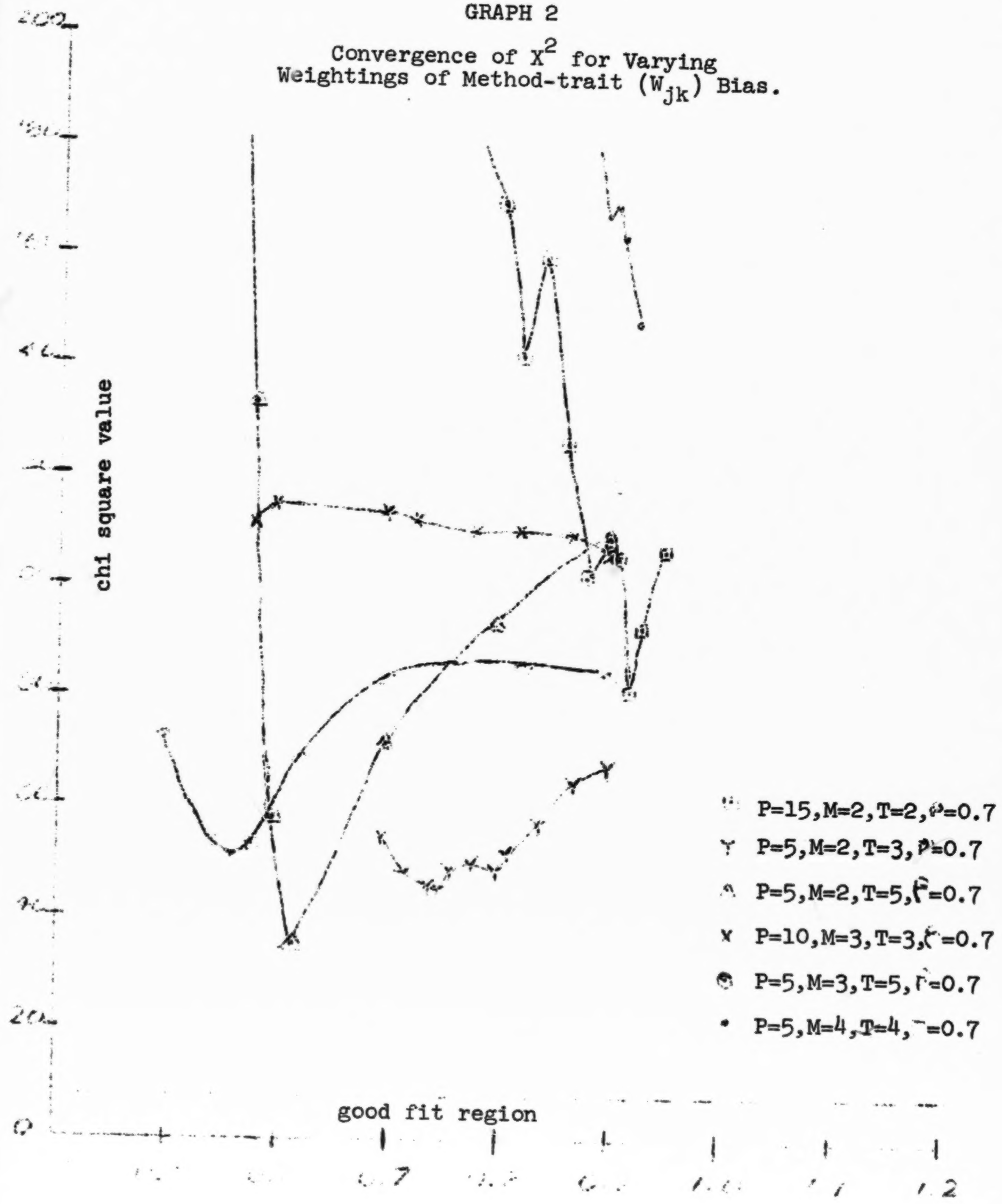


TABLE 10

Summary of the Weightings of Method (b_j) and Method-Trait (W_{jk}) Which Minimize Chi Square Values Best

M	T	P	— Emp.	— Theor.	b	W	X^2
2	2	5		0.7	8/100	92/100	*
2	2	5		0.9	"	"	*
2	2	15	0.67	0.7	"	"	81.3
2	2	15	0.87	0.9	"	"	99
2	2	25	0.68	0.7	"	"	32.7
2	2	25	0.87	0.9	"	"	42.3
2	3	5	0.73	0.7	1/4	3/4	48.5
2	3	10	0.47	0.3	"	"	74.5
2	3	10	0.71	0.7	"	"	75.1
2	3	10	0.91	0.9	"	"	67.1
2	3	20	0.71	0.7	"	"	88.2
2	4	5	0.71	0.7	38/100	62/100	28.5
2	5	5	0.70	0.7	38/100	62/100	36
3	3	10	-	-	Not clear		-
3	4	5	0.68	0.7	1/3	2/3	10.4
3	4	5	0.88	0.9	"	"	31.9
3	4	15	0.71	0.7	"	"	12.98
3	4	15	0.90	0.9	"	"	9.96
3	4	25	0.71	0.7	"	"	17.05
3	4	30	0.90	0.9	"	"	18.7
3	5	5	0.69	0.7	44/100	56/100	53.9
4	4	5	-	-	Not clear		-
4	5	5	-	-	Not clear		-

* > 100 neg. F's

Zyzanski's F_{PT} Statistic

For every case in which Zyzanski's F_{PT} statistic was generated more than 100 negative F values resulted. One hundred such values were sufficient to terminate the computer program. Empirical distributions which were terminated for this reason are not good approximations of theoretical F distributions.

Summary of Results

Computer programs were developed which successfully generated the F_{PT} statistics (for small sample sizes) of Stanley and Zyzanski which had been developed to determine validity by multitrait-multimethod matrices. Stanley's statistic was not robust under varying combinations of method (b_j) and method-trait (W_{jk}) bias but it was possible to prescribe conditions where this statistic would be useful. Zyzanski's statistic did not ever approximate a theoretical F statistic.

PART II LOGICAL ANALYSIS

I. Introduction

Measuring individual differences, we tend to think, ought to meet some criteria or other. This opinion seems to be bolstered by the belief that if we engage in an enterprise or activity, there is some right way (perhaps several right ways) of doing what we intend. For example, counting the hairs on one's forearm is not thought to be the right way to discover a personality trait like intelligence or sense of humor. This sounds rediculous, but we must remember that, with some people, the lines on the palm of one's hand can be used to discover personality traits, as well as numerous other items of interest.

The problem is to determine at least one right way of measuring traits. So the question arises; what is to count as a good test, one which we can set store in. This question might draw as response a list of tests which are considered as worthy examples of what a good test is. Like Socrates, seeking the meaning of "good," we must turn these aside and ask, "What is it in virute of which a test is good or in virtue of which the results are noteworthy?"

This question can be answered in several ways. To cut the philosophical discussion short (however dangerous and prejudicial to clarity), we can say we are in search of a definition of "good test" or that we want to know what it means to be a good test. The fact that someone presents a test on the market, as all agree, does not guarantee the worth of that test. Yet, there have been

(NOTE: For all references in the Logical Analysis refer to notes in Reference section).

few efforts to really investigate the criteria which must be met for calling a test "good." At times one gets the impression that if a test can be presented decked out with impressive statistical correlations, with charts, graphs, matrices, numbers, etc., it lays claim to being called "good." However, the gypsy who is adept in palmistry could employ some of these very same techniques; yet somehow, we remain loath to accept her conclusions as reliable and valid.*

It is this problem to which Campbell and Fiske[1] are addressing themselves: What principles can we employ in sorting out valid from invalid tests?

Campbell and Fiske's article has been praised as raising some crucial problems, and we acknowledge their contribution in stirring interest in this important area. Our study is aimed at clarifying and organizing their ideas, and, in general, furthering the work they have begun. The Campbell-Fiske approach, we feel, could be looked at from two points of view. The first point of view might be seen as that of practical rules with the aid of which one can effectively tell that the results of the test are of some worth. The second point of view is the examination of why the rules are indeed "desiderata," if not necessities.

---

* Yet as we shall see from our discussion below, the gypsy's method could be "reliable" in the technical sense of yielding similar results in the test-retest run. Suppose our gypsy counts the lines on my palm (say, four longish lines) and concludes that I am a rake. I return an hour later and present my hand (with its four longish lines) and she again flatters me by calling me a rake. Her diagnosis is "reliable." (See pp. 56 ff. below).

This latter aspect is the most important, since it would reveal the rationale behind the rules and would justify our acceptance of the four Campbell-Fiske criteria. This theoretical, as opposed to the practical, aspect of their work must be studied, therefore, before one undertakes the task of judging particular test results in the light of these criteria. In short, we want to know if the criteria are good ones.

For example, the actual values on the matrix are checked against the practical rules mentioned above. By appeal to these practical rules, the values are judged to be "reliable" and/or "valid." But the practical rules, in turn, must be justified by an appeal to the necessity, utility or desirability of the concepts which underly them. It is this latter task with which we are now occupying ourselves.

To what do the authors appeal in order to justify their criteria? One could propose various justifications. For example, we could offer an _a priori_ one. That is, we could analyze the _concepts_ we have of test, of method, of limit, _etc_., and try to show that, given our understanding of these terms, certain other things are entailed logically, necessarily. This sort of justification, we feel, would be the strongest sort possible. Necessary truths are hard to come by, however, so we may have little success in such a venture. We will, however, offer a tentative analysis of the criteria and try to determine whether or not the criteria are entailed by the notion of test, _etc_.

If no satisfactory _a priori_ justification can be discovered, the authors can very well appeal to other sorts of justification:

to the desirability of these criteria, to the utility, etc. If so, then the criteria might be seen as normative expressions of what are a posteriori generalizations. As such, the criteria are expressions based on contingent factors and may well have to be revamped and revised in the light of further evidence and experience. The status of such "contingent criteria" is obviously inferior to that of "necessary criteria."

All of these remarks, of course, appear to be highly speculative and abstract. This we do not deny. The point is that such an examination of the foundations of testing is much in need, and few people have busied themselves with these deeper problems. People who deny the value of this sort of study must be prepared also to be inconsistent, saying that we must make sure our tests are valid, but we need not worry whether our criteria for ascertaining validity are indeed correct.

This paper is an effort to obviate the problems which might arise from uncritical acceptance of test results and uncritical acceptance of norms to judge those results. Our approach will follow the lines of a conceptual analysis in an effort to ascertain what criteria are a priori and necessary for test results to be called valid and reliable. That is, our analysis will be a logical, not statistical, analysis.

Unless the criteria presented by Campbell and Fiske require some a posteriori justification, we can hope to discover that the criteria rest on some self-evident and intuitively grasped notions of personality-trait testing.

II. A "Good" Test

We have mentioned that our analysis would depart from our concept of a good test in order to uncover what criteria are entailed by such a concept. That is, if we intend to say that the very notion of a good test demands that certain criteria must be met, then the examination of the meaning of "good test" ought to reveal what criteria are required. The justification of the criteria would be that such criteria are entailed by, or follow necessarily from, the prior notion of testing.

I suppose we could proceed by saying that a test which does what we intend it to do is a good test. So we must be clear about the aim of testing and measuring personality traits. Most simply and starkly stated, the aim of personality trait testing is to discover the presence or absence of a trait and to ascertain to what degree the trait is present. This overarching fact - that such a test is an instrument aimed at discriminating properties - must be distinguished from the secondary aims such as using test results for the purpose of hiring, firing, etc.

At this common-sense, non-technical level, it is safe to say that any test which really discovers the presence and degree of the trait it is designed to measure is a good test. We also tend to speak of such a test as "valid" and "reliable," where "valid" is used interchangeably with good," and so is "reliable." We can easily imagine a frustrated admissions officer inquiring whether a certain test is indeed a good guage for selecting graduate students. His assistant, convinced that the test does pick out success-bound students, might reply in a number of ways, all of which, he might feel, amount to the same thing:

(1) Yes, the test is a good one.
(2) Yes, the test is valid.
(3) Yes, the test is reliable.
(4) Yes, the test is trustworthy.

All of these statements could be taken as saying, "Yes, the test does successfully discover the kind of student we are looking for."

This readiness we have to conflate the meanings of various words can be bewailed, but such lugubrious behavior is beside the point. What is important is to distinguish precisely what we do mean by the various terms. It is clear that these words cannot be simply interchanged in all contexts. For example, we can readily think of a test which may be "reliable" and "valid" in some technical sense (or even in ordinary use), but which is no good for our purposes at a certain time. In one sense it is a good test for people interested in a trait (T_1), but it is not good (= useful) for someone not interested in Trait T_1.

I do not think it is wholly inacurrate to say that most people might agree with our simple-minded "definition" of a good test, given above. But the next move is to equate "good" with "possessing reliability and/or validity." Certainly, a good test ought to be reliable and valid. But good need not mean "reliable and valid." And, further, it is not clear that "reliable" and "valid" retain their original "ordinary use" meanings when we go farther into the realm of testing. If there are certain contexts, as in the case of our admissions officer, where these words can be used interchangeably, then there are just as certainly some situations where "good," for example, cannot be substituted for "reliable." We shall see that this is so for Campbell and Fiske's technical use of "reliable," (compare Cronbach, Essentials of Psychological Testing, 1960, on reliability. pg. 126 ff.).

It may be quite possible that someone would set-up some criteria for validity and reliability, only to find out that, even when these criteria are met, we hesitate to call it a "good" test.

All this amounts to a warning that we must be careful not to use words in such a way that they trade on other senses or meanings of the same word. We must be careful, for example, to distinguish "reliable = yielding consistent results" from "reliable = trustworthy." And if we do, at the common sense level, demand that a good test be "reliable" (= trustworthy), then let us be certain that "reliable" (= yielding consistent results) is not taken as its substitute. Unfortunately, some of the literature, at least, suffers from a dismal failure to effectively define these crucial terms. We have tried to show this thus far in our examples using the word "reliable." Let us comment briefly on the plight of "valid."

III. The Meaning of "Valid"

To say that test-results* are valid involves one in difficulties comparable to those which we encountered when discussing "reliable." Just what is meant when one says that test results are valid? We must be careful not to conflate this use of "valid" with some other possibly more familiar use of "valid." For example, in deductive logic, one can say that a conclusion is valid, if one has arrived at that conclusion in accordance with a rule of inference. To say that test results are valid however, does not seem to mean the same thing.

The problem seems to be that the notion of validity, even though discussed at length in books on testing (e.g. Cronbach, Essentials of Psychological Testing, 1960, chap. 5), still needs clarification. Different kinds of validity are postulated, as in Cronback, pp. 103 ff:

(1) predictive validity,
(2) concurrent validity
(3) content validity,
(4) construct validity,

Campbell and Fiske speak of

(5) convergent validity (abbreviation CV) and
(6) discriminant validity (abbreviation DV) (Campbell and

---

* The criteria for judging whether a test and test-results are valid can be discussed together. We can say that a test is valid if the results which it yields are satisfactory (valid). Then we can concentrate on the results only and try to determine the criteria whereby we can judge the results.

Fiske, pp. 81-83), and they go on to ask only for

(7) relative validity,

and not

(8) absolute validity.*

Campbell and Fiske say that their discussion of convergent validation touches all but content validity.[2] Furthermore, the Campbell-Fiske notion of validity sees validity as eventually shading into reliability.**

The picture is further complicated by the fact that for a test or test results to be counted (9) valid (simpliciter?) by Campbell and Fiske, the test or results must have (5) convergent validity and (6) discriminant validity. Criteria are offered in order to distinguish whether a test has either (5), or (6), or both (5) and (6). If criterion I is met, then the results are convergently valid (5); if criteria II through IV are met (6), then the results are discriminantly valid. And it seems to be their opinion that we are in a position to call a test valid simpliciter unless both CV and DV are present.

The entire point of these remarks is to show that although the word "valid" may creep innocently into a discussion and seems to demand acceptance as some sort of intuitively grasped, clear-cut and well-defined term there are absolutely no grounds for assuming that this is the case. And there is no reason to suspect that the ordinary language use of "valid" can serve as an overarching explanation of these various uses. Uses (1) through (6) clearly are put forward as some sort of technical uses. The others, (7) through (9), might perhaps be "ordinary uses" of the word, but the burden of proof is on those who care to hold such a position. Our recommendations thus far are:

---

* "In practice, perhaps all that can be hoped for is evidence for relative validity, that is, for common variance specific to a trait, above and beyond shared method variance."[3]

** See the remark, "Independence is, of course, a matter of degree and in this sense, reliability and validity can be seen as regions on a continuim."[3]

1.) That the notion of validity in testing be thoroughly examined and defined, so that it can become clear if (and how) such a notion can be related to our common sense intuitions about validity and to the technical notion of logical validity:
2.) That extreme care be taken in distinguishing our common sense uses from technical uses of words.

The literature contains discussions of "valid tests" and "reliable tests," but these notions are not always directly and clearly related to the notion of "good" or "valid" test with which we begin our inquiries. Equivocation can easily occur in such a situation. Many things seem to be considered as intuitively clear: the notions of test, method and trait; the aims of testing; and some of the properties of tests like goodness, validity and reliability. Our laconic comment is: Are they so clear?

IV. Campbell and Fiske: Towards a Definition of a Good Test

Let us now concentrate on Campbell and Fiske's approach to see how they try to clarify the concept of test-validity. Campbell and Fiske are trying to present some criteria whereby we can ascertain whether test results are indeed valid. How they use the word "valid" will emerge as we discuss the criteria which they propose. It will be assumed in this paper that the reader is acquainted with Campbell and Fiske's article cited above, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," in the Psychological Bulletin 56 (March, 1959), 81-105.

The criteria are presented as "common sense desiderata." Presumably, they follow from what we think a good test ought to be. The kind of test being discussed here is the personality-trait test, and its aim could be seen (1) as determining whether or not a certain trait (*e.g.*, intelligence, leadership, *etc*.) is posessed by (or present in) a person, and (2) as further determining to what degree the trait is present. This will require what statisticians call nominal and ordinal scales, and at times even interval scales.

These tests, then, aim at discerning which people have trait T (or property P), and are constructed in such a way as to screen out possessors of T from other members of the population or sample and, at times, to rank the possessors of T.

One of the problems which Campbell and Fiske wrestle with hinges on our viewing such a test as a trait-method unit.

> Each test or task employed for measurement purposes is a *trait-method unit*, a union of a particular trait content with measurement procedures not specific to that content. The systematic variance among test scores can be due to responses to the measurement features as well as responses to the trait content.[4]

This ushers in the problem of method-variance, and influences, we believe, the choice of criteria which Campbell and Fiske end up with. It is their belief that the result one arrives at when measuring a trait is *not* due simply to the trait and the amount or degree of the trait present. On the contrary, the claim goes, the method which one employs introduces unwanted effects which distort the final report on the trait which the test is intended to yield.

> In any given psychological measuring device, there are certain features or stimuli introduced specifically to represent the trait that it is intended to measure. There are other features which are characteristic of the method being employed, features which could also be present in efforts to measure other quite different traits. The test, or rating scale, or other device, almost inevitably elicits systematic variance in response due to both groups of features. To the extent that irrelevant method variance contributes to the scores obtained, these scores are invalid.[5]

The reason for postulating method variance as an explanatory factor arises from the fact that some tests, when administered for the purpose of measuring putatively independent traits, tend to yield the same or similar results for each and every trait. These questions then arise: (1) Should these various traits not show up in varying degrees? (2) And ought not a particular method be better at uncovering a particular trait, rather than a whole series of traits? We shall return to these questions. But perhaps the best way of posing the tester's dilemma is: Can such a test be good? There is a straightforward way of taking this question as a way of saying that the test is just plain useless and that we ought to jettison the test for another. But there is also the approach which says that there is trouble with this test which is due to method factor. If one could ascertain how much method variance or apparatus variance entered into our results, we could determine the amount of the trait present.

Some of the possibilities which arise when we have a test which yields the same result for each and every trait are:

(1) the test is worthless, in the same sense that counting the hairs on my arm is worthless when determining my I.Q.

(2) the traits are in fact one and the same or are not independent.

(3) the traits ARE all present to an equal degree (although many tests seem to assume this is not so, it is logic-

ally possible that this state of affairs obtain, provided the traits are not mutually exclusive by definition).

The fourth alternative seems to enter with the notion of method variance.

(4) Method variance, which is allegedly explanatory of a part of every test result, is very high. This seems to amount to more than is said in (1) above, since (4) implies, it seems, that the test can be treated in ways which may still make it useable. To the non-expert this appears at times to be an unwillingness to grant that there can be blatantly and totally inappropriate tests.

Campbell and Fiske would, it seems, condemn the sort of defective test under question as useless or undesirable. But there seems also to be the implication that a test can be all right if its method variance can be determined and if the methods have certain properties like convergence and discrimination. That is, this method-variance which "invalidates" one's results can be detected, and the overall validity or validity sampliciter of a test can be determined if one has results which are convergently valid and discriminantly valid. This bifurcate-validity can be ascertained, however, only if one employs a multi-trait and multi-method approach.

One thing is clear, however: Campbell and Fiske are offering some definite criteria whereby we can judge the worth of a test. Presumably, if a test meets their criteria, the test is good.

In the light of Campbell and Fiske's criteria, there arises a need for a multitrait-multimethod apprcach. That is, more than one trait and more than one method are required if we are to be able to KNOW WHETHER THE TEST IS GOOD (= RELIABLE AND VALID). The use of a multitrait-multimethod matrix can be used to portray reliability and validity; and failure of the matrix to meet the form proposed criteria would seem to be explained by the fact that the matrix is only apparently multitrait - multimethod. That is, a defective matrix might be shown to be (i.e., reduced to): (a) a monotrait - monomethod matrix (which would not reveal validity), or (b) a monotrait - multimethod matrix (which would not evidence "discriminant validity"), or (c) a multitrait monomethod matrix (which would not display convergent or discriminant validity).

The fact that matrices of the sort (a) through (c) do not permit one to ascertain the validity of the array of values in the matrix prompts Campbell and Fiske to stipulate the multitrait-multimethod matrix as necessary for revelation of validity. This is borne out be statements like the following:

> ...The clear cut demonstration of the presence of method variance requires both several traits and several methods. Otherwise, high correlations between tests might be explained as due either to basic trait similarity or to shared method variance. In the multitrait-

> multimethod matrix, the presence of method variance is indicated by the difference in level of correlation between the parallel values of the monomethod block and the heteromethod blocks, assuming comparable reliabilities among all tests.[6]*

Since a multitrait-multimethod matrix is designed to reveal reliability and validity, we might assume that it will reveal whether a test is good or not. One could fairly, I think, take reliability and validity as sufficient criteria for calling a test good.

$$\ulcorner G(T) \equiv R(T) \wedge V(T) \urcorner .$$

(A) Multitrait-Multimethod Approach and Reliability

To ascertain whether a test is good, then, we can begin by asking, "Are the results reliable?" To answer this question, one must set out the criterion of reliability. For Campbell and Fiske, reliability is present if the results of a given test or method, M_1, which is designed to measure a given trait, T_1, correlate at 1.0 (ideally) with the results of another test M_2 for T_2, where $M_1 = M_2$ and $T_1 = T_2$. In actual fact the

* Note also: "Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods."[7]

correlation *approximates* 1.0. The rationale behind this definition of reliability is that the testing method designed to measure a particular trait should yield identical results when reapplied. The situation thus ideally described, however, becomes more complex in concrete instances due to the change of circumstances, test-sophistication, etc.

Thus "Reliability" seems to rest on the notion that a test should yield nearly the same results when administered two (or more) times to the same person under maximally similar circumstances. Reliability, in this technical sense therefore does *not* mean that the test method is a reliable guage of whether or not a person does have a trait or not. Indeed, the test method may be what one might call an unreliable guide for judging whether or not Jones is intelligent. *What* the method says is irrelevant to this definition: *it is only important that the method keep yielding maximally similar results* whose correlation approaches 1.0. "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods."[8]

Obviously, we are not satisfied with this sort of reliability alone, since one can, and ought to, raise the question: Are these test results truly indicative of the degree to which a trait is present in a person? How can we know? Perhaps the data, though reliable (in the sense given above), is wrong - i.e., suppose we keep getting the SAME (∴ "reliable") DECEPTIVE RESULTS. The gypsy, for example, can arrive at the same results each time she applies her palmistry methods - which are "reliable" in this sense.

Ideally, we would check for the same trait by a totally independent method.* We must assume that the two methods are _effective_, _i.e._, that they really work. It is logically possible for someone to check a single trait T by a finite number of independent methods $M_1, \ldots, M_n$, and in fact employ a non-independent or a _defective_ method every time. M_{n+1} would perhaps be a good one, but the tester gives up before reaching it, and ends up thinking his data are "valid." Assuming that these two methods

---

* Note: The notions of "independent method" and "effective method" are important and ought to be examined thoroughly. The notion of independence is central to the entire discussion of the multitrait-multimethod venture, since by "multi-X," the authors are speaking of two-or-more-independent-x's, either methods or traits. But the concept of independence is not defined. The authors do not say that independence is to be an intuitively grasped term, but they do indeed proceed as though such were the case. The problem is that independence is not intuitively clear. Even if one trys some ordinary language renderings of this technical term "independence," one is not much enlightened: _e.g._ To say x_1 and x_2 are independent means they are not the same, not identical....One could go on like this, but with little profit. What is required is a clear definition of independence. Or, if such is impossible due to the fact that this concept is primitive and is the concept in terms of which other concepts are defined, then there ought at least to be some further analysis of what is entailed by independence. _A propos_ our project, this would be helpful in explaining why the four criteria of Campbell and Fiske make the demands they presently make. Since convergent validity, for example, is defined in terms of independent methods converging on the same trait, it would be helpful to know what is meant by "independent." Indeed, the whole multitrait-multimethod is composed of a complex of independent methods and independent traits.

M_1 and M_2 are independent and also do effectively measure T_1, then we can expect that there will be some degree (hopefully a high degree) of correlation between the results of M_1T_1 and M_2T_1. Some such demand is necessary to augment the above definition of "reliability."

(B) Validity and the Matrix

This brings us into the discussion of validity and its problems. "Reliability," as such, guarantees us nothing or, at best, very little. Our common sense requirement, mentioned earlier (that a good test be one on which we can rely, and whose results are trustworthy, and which really does measure the desired trait), is neither fulfilled nor guaranteed by such a definition of reliability. Assuming a test does measure the same thing twice, however, we cannot deny that the results ought to be similar as long as the thing measured is postulated as remaining the same.

Hence the demand for validity, and the demand for "convergent" validity made above. And hence the need for a multimethod approach. The convergence of methods is meant to insure against the danger inherent in the use of only one (possibly deceptive) method.

(1) Convergent Validity: Criterion I

The first criterion, which assumes use of convergent validity, thus makes its appearance.

> In the first place, the entries in the validity diagonal should be significantly different from zero and sufficiently large to encourage further examination of validity.[9]

Accordingly, the values of a validity diagonal (= the monotrait-heteromethod diagonal) must be greater than 0 and sufficiently large.

$$V_c(T) = (\left[C(m_1t_1),(m_2t_1)\right] \wedge (C\ 0) \wedge (C=N)\ )^*$$

The motivation for this criterion is the belief that two independent methods designed to test the very same trait ought to yield a high correlation - they ought to yield similar results, where "similar" is left vague, hazy and undefined. A problem here is to say that the values of such correlations should be "sufficiently large" leaves us desirous of further clarification. The authors may want to say that the criterion of "largeness" is a function of a particular matter under study. This ploy would allow the notion of "largeness" to take on meaning relative to a given series of traits, methods and circumstances.

Probably - almost certainly - the authors want built into this criterion the idea that the methods employed are independent and effective. This helps obviate the problem of convergence of defective or poor methods (thus making the result a monomethod - monotrait correlation, which is a "reliability" result). Once

*(Note that "T" now stands for Test, while the lower-case "t" stands for "trait," and "m" stands for "method," C for "Correlates with," "V_c" for "convergently valid," "V_d" for "discriminantly valid," and "V_{cs}" for "common-sensically valid."

we grant that we have two such effective methods for measuring T_1, we can safely, indeed trivially, conclude that the convergence of these methods ought to be greater then 0 and fairly high. To conclude otherwise would land us in self-contradiction.

Perhaps the importance and soundness of this criterion can be seen in situations where it is NOT met. If two methods, putatively independent, are measuring the same trait, then as effective methods, they ought to reveal whether or not the trait is present. If the correlation is zero, then one suspects that the two methods are not designed as effective measures of T_1, but perhaps are after different traits. The methods, if they do not converge, do not serve to check one another out a propos the same trait --- quite obviously.

If the methods correlate at some value other than zero, but not very high, then it seems odd to say both methods are effective---since they draw different conclusions about a trait they are both supposed to measure accurately.

But what if the correlation is more than "sufficiently large?" Suppose the correlation is as large as possible, viz., + 1.0? In this case, then, we may have:

(a) really a monomethod-monotrait situation, and the two tests are really not independent, so compose a "reliability" test, not a "convergent validity" test,

or (b) there is no method factor present. That is, two methods could yield exactly the same results about exactly the same property, so that "method variance" ---as a contributory factor to the results - seems non-existent (or incapable of detection). This can certainly happen in testing mathematical expressions. Two results can regularly correlate at +1.0.

The whole question of method variance then rises up like a specter to haunt our discussion.

Method factor may not be a plague which besets all trait-measuring. It might well be confined to the kinds of personality-trait tests we are considering. If so, then the "trait-method unit" doctrine can be seen as a postulate for work in this field. But it cannot claim to escape challenge, as though the "trait-method" combination followed analytically from the definition of "test."

It must be pointed out that simply because we have what we call "a method," we are by no means justified in assuming that two such "methods" will in all cases give us valid data, in a favored sense of valid (= trustworthy, sound, etc.). It is quite possible that M_1 and M_2 (where $M_1 \neq M_2$) could both be unsound, poor, deceptive methods of measuring a trait. The fact that we call a thing a method does not entail that it is a good, effective method. Otherwise we would never be able to speak of poor or bad methods, since the word "method" would mean "good method" and we would be talking nonsense about "bad(good)methods." Campbell and Fiske, of course, say nothing contrary to what we are saying here. But this is an underlying presupposition of their criteria. The problem of "method factor" leads one to expect that for any M_i, M_i will to some extent influence the measurements. However, it is possible that a certain method, M_c, be entirely useless. To use the current terminology, it seems possible that the results from a test be entirely due to method factor. This even seems possible on the "trait-method unit" view. One could say that as the trait factor decreases, the method factor increases. And if, as seems possible, a method be entirely responsible for the results, then the method is 100% useless. That there can be useless and totally inappropriate tests where method factor seems to play no part - in mathematics we can construct two independent tests for a certain property. The validity correlation can be 1.0 (equal to a reliability correlation) and there is no way to determine if there is such a thing here as "method factor."

The core of the problem of method variance seems to be in factor analysis, where the method is seen as always influencing the results. In our mathematical cases, however, it is difficult to see what could be meant by method factor. It is hard to conceive how the method of determining the algebraic sign of the root(s) of a polynominal could "influence" the test result.

Perhaps the problem of method variance could be subsumed under some of the main problems of philosophy like the problem of "seeing as" (e.g., as discussed by Wittgenstein), or the problem as presented by Kantian-minded philosophers of science. The means of observation cannot be ignored, and it is not our intention to look down upon any efforts to come to grips with the contribution to our knowledge made by our means of observation. What we do want to say is that the criteria which was suggested on the assumption that method-factor is always involved must be given a critical going-over. We ought to question the assumption, and we ought also to inquire whether or not the criteria follow of necessity from our ideas about testing, or are dictated by other considerations, e.g., experience and utility.

If the criteria are based on experience, then they ought to be susceptible to revision again and again in the light of experience. The big danger is that if the criteria become entrenched, then they may be used to rule out of court certain results which do not meet the criteria as presently stated, in the light of which results the criteria ought to be revised.

Part of the solution to the problem seems to lie in examining the view that the test is a "trait-method unit." (See above pp. 51 ff.). The whole business of method variance as stated above lacks cogency, it seems. Simply because M_1 and M_2 share certain features in common,[10] it does not follow that these common features combined with a single method's unique features will draw some responses appropriate to the unique features and some appropriate to the common features. The method's having some elements in common with another method entails nothing. Why is it not possible for E_1 to combine with E_2 in a way which yields a totally unique "molecular" structure, as H_2O yields a molecule of water---though, obviously H and O are held in common by numerous other molecules.

(2) Discriminant Validity: Criteria II-IV

Criteria II-IV provide the means of determining "discriminant validity." (= DV) This DV demands at least two independent traits and two methods. When M_1T_1 and M_2T_2 correlate as highly as M_1T_1 and M_2T_1, then there is no discriminant validity, nor when M_1T_1 and M_1T_2 correlate higher than M_1T_1 and M_2T_1 (where M_1 and M_2 are methods designed to get at T_1 and T_2, respectively, independent of any other traits).

The reason for postulating the need for discriminant validity is the idea that to verify the existence of (and to measure) distinct traits requires distinct, specially constructed methods. Campbell and Fiske explain their reasons for expecting DV in a test in passages like the following:

> "When a dimension of personality is hypothesized, when a construct is proposed, the proponent invariably has in mind distinctions between the new dimension and other constructs already in use. One cannot define without implying distinctions and the verification of these distinctions is an important part of the validational process."[11]

However, it is logically possible for one method to determine very accurately the existence (or degree) of two or more traits. It is possible to conceive that wherever there is T_1, there also is T_2, where T_1 and T_2 are independent, but universally and contingently accompany one another, but are neither logically nor causally related.

CRITERION II

> "Second, a validity diagonal value should be higher than the values lying in its column and row in the hetero-trait -heteromethod triangles. That is, a validity value for a variable should be higher than the correlations obtained between that variable and any other variable having neither trait nor method in common.[12]

Granted that the methods and traits are independent,

$$C(m_1t_1, m_2t_1) > C(m_1t_1, m_2t_2).$$

The assumption, of course, is that where all the factors differ, there should be a lower correlation. The general assumption is that there is an inverse ratio between the amount of difference between factors and the correlation of results. Hence, there seems to be no contradiction in denying this apparent demand made by DV. The assumption, "where the trait differs, there also the method should differ," needs deeper scrutiny. At present there seems to be no logical necessity for it. But let us look at the criteria for DV in order to understand its requirements as well as possible.

A few statements can be made at this juncture:

(1) two results, M_1T_1 and M_2T_1 could correlate highly, as we saw previously when discussing convergent validity.

(2) Two results M_1T_1 and M_2T_1 need not correlate highly, if neither are effective methods, or even if one is a defective method.

(3) two results M_1T_1 and M_1T_2 could correlate highly, though they need not, as was just said on page 66

(4) It is possible that we test for T_1 by means of M_1 and T_2 by means of M_2. Again, we see that it is logically possible due to a constant conjunction, to use Hume's language, to have a high correlation, since (a) M_1 and M_2 may be effective for their respective traits and, (b) T_1 and T_2 may be constantly (though not of necessity) conjoined.

Campbell and Fiske's criteria deal mainly in terms of correlations. These criteria specify that certain results of testing ought to correlate in a certain way with some other results. But our examination reveals that one can deny the necessity of such criteria or requirements without landing oneself in a contradiction. This will emerge again when we discuss criteria III and IV in what follows. THIS IS NOT TO SAY THAT THE CRITERIA CANNOT BE GROUNDED ON PRINCIPLES OF EXPERIENCE, SUCH AS UTILITY. BUT IT IS TO SAY THAT THE CRITERIA FOR DISCRIMINANT VALIDITY ARE DEMANDS PLACED ON TESTING THAT ARE NOT IMPOSED BY LOGICAL NECESSITY.

CRITERION III

> "A third common sense desideratum is that a variable correlates higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method."[13]

They go on to add:

> "For a given variable, this involves comparing its values in the validity diagonals with its values in the heterotrait-monomethod triangles."[14]

But this criterion is not always met, a feature which "is probably typical of the usual case in individual differences research." Even Campbell and Fiske's synthetic matrix fails to meet this criterion satisfactorily.

The problem with this common-sense desideratum is that we have difficulty in seeing why it *must* be desired. That many people *do* desire it proves little, if anything, at this stage. They may well be desiring something quite useless, or impossible. One thing that does emerge is that what they desire is not *necessary* in the sense of logically necessary. It is quite possible that one method reveal two properties which are independent (as in our constantly conjoined traits cited above).

It also seems that it is possible for the correlation of M_1T_1 and M_1T_2 to be higher than $M_1T_1 \wedge M_2T_1$, since M_2 might well be a far poorer (a less adequate) instrument than M_1 for discovering the presence of (and/or amount of) T_1: Although M_1 might be well adapted in this fashion to measure T_1 and T_2.

$$\Diamond (M_1T_1 \wedge M_1T_2 = X)$$

$$\Diamond (M_1T_1 \wedge M_2T_1 < X)^*$$

---

* Note: The sign diamond $\Diamond$ has its traditional logical modal significance often interpreted as "It is possible that...." "X" is considered here as some high correlation considered trustworthy.

The entire problem seems to be a question of distinguishing what must be (or ought to be) from what generally does happen to be (even though it happens to be in many "useful" and "good" tests). There seems to be no a priori need for Criterion III. If a justification is to be given a posteriori, then cases must be adduced (a) where it has been met, and (b) where the fact that it has been met is significant or important. If this is not done, one can keep the "criterion" in mind to see if enough evidence arises to validate this "criterion," but they cannot use this "criterion" as a norm against which test data are held for judgment.

CRITERION IV

> A fourth desideratum is that the same pattern of trait inter-relationship be shown in all the heterotrait triangles of both the monomethod and heteromethod blocks.[15]

What this seems to be requesting, prima facie, is that a trait show a regular pattern of relationships when that trait is measured by the same or different methods.

This seems to assume that if a trait is present, it will reveal itself in a constant fashion as being related thus-and-so to any other trait which is present. Thus, the correlations on Campbell and Fiske's Synthetic Matrix maintain a certain pattern of values in the heterotrait triangle.

In order to have such a criterion hold, we seem to be obliged to stipulate that certain presuppositions hold. These presuppositions would be the notions found in Criteria II and III or some other set of ideas about methods and traits, which entails that such a pattern of relationships hold. Given a group of truly effective methods and a group of traits, the methods should reveal the relationships which actually obtain among the traits. If contradictory results are obtained, then there is reason to inquire into the efficacy of the methods. Criterion IV, if it demands only this, is all right. But it seems to be saying much more than this.

One sure test for this criterion would be the construction of a matrix which was based on logically sound grounds, but which has at least two distinguishable patterns. Such a counter-example would put an end to any discussion of the logical necessity of this criterion, unless it is interpreted in the trivial sense explained above.

V. Conclusion

Many minor points might be mentioned as a result of our investigation, also some remarks of a highly general and highly important nature. For example, there is a clear need for an effort to get below the work-a-day testing procedures and problems to try to see why a test is good, or why not. Campbell and Fiske have made an effort to delve into the rules which govern good testing, and the issue needs further work and critical scrutiny.

Also, there are a number of crucial and basic, yet unsatisfactorily defined, concepts which are employed in methodological discussions.

But specific to our discussion, I feel there are two points which deserve special consideration. First, the status or foundations of criteria must be determined before we can judge their worth. And surely, criteria no more deserve to escape critical examination then anything else. If criteria are seen as custodians of good method and procedure, we must make sure we do not get tongue-lashed by the Roman satirist Juvenal: _Quis custodiet ipsos custodes_? This has been our task - to avoid being uncritical of the standards employed. The criteria presented by Campbell and Fiske seem to make demands which go beyond the logic of the concepts involved. If it is possible to have a good test without all these criteria (and it is logically possible), then we cannot blindly follow such rules and exclude tests and results which might be trustworthy, though not canonized by our four criteria. This would be undesirable, and perhaps wasteful.

As we acknowledged earlier, the criteria may have justification other than logical necessity. Economy, speed, _etc_., may dictate the employment of such criteria. But in that case, we cannot be smug about the sentence we pass on "invalid" test results. Perhaps further experience will reveal that our criteria need revising in light of recent discoveries. Some of the results ruled out of court by these criteria may well be worthy of consideration and serve as the basis for revision of the rules. This caution about handing down rulings on tests is in place once we see that criteria cannot stand without appealing to experience for justification. Experience can alter one's views, as well as justify them.

Secondly, the entire approach which we have taken toward the examination of the criteria may be fraught with difficulty.

We said that if a method was "effective," that it was successful in measuring a trait. Such a procedure seems tantamount to saying, "If the method is effective, it gives valid results." If so, then we must re-examine our work to be sure we have not been unfair nor inaccurate. For otherwise it would seem that one must presuppose validity in order to account for it. This would end us up in a vicious circle.

It may be possible that Campbell and Fiske's criteria do rest on circularity, but it may well be that my account forces it into a vicious circle. The issue deserves consideration. At present it seems that only criterion I definitely holds, and possibly criterion IV, on a trivial interpretation. In both cases, however, we had to invoke the notion of effectiveness in methodology (valid methodology?) to arrive at acceptable interpretations. If so, then these criteria, which are meant to lead us to an understanding of validity, presuppose that we already understand this concept. And to discover whether the results are valid, we seem forced into granting that the methods must be valid. Then it would follow that the results are valid.... And so on. The vicious circle rolls on and on. If the interpretations I put on the criteria lead to this situation, then the criteria SO INTERPRETED WOULD BE USELESS.

The problem might be more clearly illustrated in the following way:

Tests are being cranked out, and we want a way to separate the good ones from the bad. One way, say Campbell and Fiske, is to check to see whether the test results conform to the four criteria discussed above. However, our critique of the criteria showed that one could very well meet these criteria, as well as have a "reliable" test, and we could still consider the test as untrustworthy and as not good from the common sense point of view. Convergent validity did hold, however, cnce we put certain explicit restrictions on it (...*if* the methods are independent, and *if* the methods are indeed *effective*.... - see page 60 ff. above). But by saying that the methods had to be *effective*, we in fact stipulated that they had to be *valid*, *trustworthy*, and *good* in the common sense fashion. But this common sense notion is what the criteria are supposed to *explain*, *not assume*. In short, the criterion to be of use, must assume the presence of the property, whose existence is uncertain as of yet. This *petitio principii*, or circular reasoning, is illustrated in textbooks on logic by examples similar to the following:

A. "I know Jones is beligerent."

B. "How do you know that?"

A. "Because Jones is bellicose."

Our present version of the problem might be illustrated in this way:

A. "The test is common sensically valid."

B. "Why?"

A. "Because it is convergently valid."

B. "Why is it convergently valid?"

A. "Because it is common sensically valid."

Effectiveness, which is a necessary condition of CV is also a sufficient condition of common sense validity. (In fact, it might be possible to define common-sense validity of _independent tests_ in such a way as to end up with the same definition as CV.*

$$(1)\quad \underset{T_1}{A}(\mathrm{Vcs}\ (T_1) \equiv \underset{T_1,T_2}{\exists}(\ (T_1 \neq T_2) \wedge E(T_1) \wedge E(T_2) \wedge C(T_1,T_2) \wedge C > 0$$

Compare this definition with that of CV.

$$(2)\quad \underset{T}{A}(\mathrm{Vc}(T)) \equiv \underset{m_1,m_2}{\exists}(m_1 \neq m_2) \wedge E(m_1) \wedge E(m_2) \wedge C(m_1,m_2) \wedge C > 0$$

Where T appears in (1), m appears in 2.

The problem is that (1) actually says more than our common sense intuition at first demands. Our common sense notion of Validity reads:

$$(3)\quad \mathrm{Vcs}(T) \equiv E(T)$$

Campbell and Fiske have been presented in our critique as offering a technical definition of validity which would be logically equivalent to (3):

$$(4)\quad \mathrm{Vcs}(T) \equiv \mathrm{Vc}(T) \wedge \mathrm{Vd}(T)$$

---

* $T_1 = T_2$ means the same as $I(T_1, T_2)$.

We can refer to (4) as the Campbell and Fiske transformation of (3). However, we feel that the criteria for discriminant validity, V_D, were neither logically necessary nor sufficient to constitute a test as Vcs, as a test could meet this criteria and still not merit the title of Vcs. Therefore, we drop V_D from (4), and arrive at:

(5) $Vcs \equiv Vc(T)$

But this is precisely what Campbell and Fiske want to avoid - the conflation of Vcs with Vc. How they will solve the problem is not our concern here. Suffice it to say here that (5) could not stand up under criticism, either, and (5) cannot be considered even as a sufficient condition for Vcs. Indeed, unless certain specific modifications are made, we cannot even consider Vc as defined by Campbell and Fiske as a necessary condition for Vcs. We, therefore, redefine Vc:

(6) $Vc(T) = (m_1 \neq m_2) \wedge E(m_1) \wedge E(m_2) \wedge C(m_1, m_2) \wedge C > 0$

which is a version of (2) above. Campbell and Fiske deny that Vc is a sufficient condition for Vcs, and this can be stated:

(7) $\sim (Vcs(T) \equiv Vc(T))$

They are not adverse to saying that Vc is a necessary condition, so that:

(8) $Vcs(T) \supset Vc(T)$

where (9) $Vc(T) \supset (m_1 \neq m_2) \wedge E(m_1) \wedge E(m_2)$

(arrived at from (6) above - that is, convergent validity requires by definition, or of necessity, that the two methods be independent and effective).

Now any test, T_i, can be one m_i or a conjunction of m_i's:

(10) $T = m_1 v m_2 v \ldots . m_n v (m_1 \wedge m_2) v (m_1\ m_2 \wedge \ldots . m_n)$

Let (11) $T_1 = m_1$

Then, substituting T_1 for m_1 in (9) above, we get:

(12) $Vc(T_1) \supset (T_1 \neq m_2) \wedge E(T_1) \wedge E(m_2)$

We then see that

(13) $Vc(T_1) \supset E(T_1)$

(which is arrived at from (12) by conjunctive simplification).

But recall (3) above, and compare (3) and (13)

(3) $Vcs(T) \equiv E(T)$

(13) $Vc(T_1) \supset E(T_1)$

If we substitute the left hand side of the equivalence in (3) for the consequent in (13) --- assuming that $T = T_1$, then we arrive at:

(14) $Vc(T_1) \supset Vcs(T_1)$

This conclusion is the one which Campbell and Fiske wish to avoid, but we seem to be lead to it if we modify Criterion I in such a way as to make it logically necessary. What we ultimately end up with is an equivalence between Vcs and Vc:

(15) $Vc(T) \equiv Vcs(T)$

(Which is arrived at from (8) and (14) - from mutual implication).

This might serve to illustrate what was referred to as the "circularity" in reasoning. That is:

(1) We say a test is Vc on the basis of Vcs; that is, Vc, to be defined requires that a test be Vcs.

(2) Then we say that a test is Vcs on the basis of the test's being Vc. But a test can be Vc only on the basis of its being Vcs. Thus, the circle.

As we said before, we may require clarification from Campbell and Fiske before we can ultimately decide the issue. We welcome correction and suggestions. Indeed, if we are to recast Campbell and Fiske's criteria in a way which can avoid the difficulties discovered in our study and this circularity, we will definitely need further study and suggestions.

DISCUSSION

This was a two part investigation. The first part was a Monte Carlo (statistical) analysis, and the second was a logical analysis of multitrait-multimethod validity. In this section Part I and Part II will be discussed separately.

Part I Monte Carlo Analysis

This part of the study succeeded in generating, for small sample sizes, empirical distributions of Stanley's F statistic for testing trait validity in multitrait-multimethod matrices. This statistic was not robust and did not remain invariant for various combinations of non-null contributions of the sources of method and method-trait bias. However, it was possible to prescribe, for most of the matrices investigated, those weightings of method and method-trait bias which would give minimal distortions of the empirical from the theoretical distribution functions.

Zyzanski's statistic, which is a correction of Stanley's, could not be generated successfully for small sample sizes without producing more than 10 per cent negative F values. Zyzanski's correction is thus inappropriate to apply for small sample sizes.

This study was limited to a scatter sampling of combinations of persons, methods, traits, and correlations because of the enormous number of calculations and the hours of computer time required. This was a limitation of the Monte Carlo Analysis and caution must be exercised in extrapotating the results. However, on the basis of the more than 150 empirical distribution functions which were generated, each with 1000 points, at a total expenditure of more than 10 hours of computer time, it is concluded that conditions can be prescribed for using Stanley's F statistic. In addition, other corrections than that of Zyzanski's might further reduce the distortions of the empirical from the theoretical distribution for the non-null contributions of method and method-trait bias.

Part II Logical Analysis

This part of the study employed the method of logical analysis to determine the soundness of the four criteria proposed by Campbell and Fiske for determining trait validity by multitrait-multimethod matrices. Our task was to determine what grounds Campbell and Fiske had for saying that their criteria _must_ be met by any good test.

Our conclusions were (1) that only criterion I could be considered a "theorem" of testing theory, and even then, _only_ after some riders had been attached, (2) that Criteria II - III seemed not to be entailed by the concepts of "test" and "validity," (3) that modification of criterion I, as we presented it, involved us in circular reasoning; and (4) that there may well be other, non-deductive ways of validating the criteria (*e.g.*, utility, convenience, *etc.*). This does not amount to a rejection of the criteria, but does implicitly make this request.

This analysis questioned whether specific tests can be validated or invalidated when the criteria offered to do this are themselves not "valid" or logically necessary. Under these conditions, applications of such criteria or principles can hardly be satisfactory.

CONCLUSIONS AND IMPLICATIONS

This investigation utilized the techniques of Monte Carlo and Logical Analyses. The logical analysis showed that there are immense difficulties which must be overcome before it is possible to give a rigorous answer to the question which asks which tests are "good" or valid. The concepts which underlie the field of testing and the logical interrelationships of these concepts are themselves not clear. Even the most "thorough" treatments of test-validity are decidedly lacking in thoroughness, logical rigor and conceptual clarity. This analysis led to the following conclusions:

(1) That only criterion I of the Campbell-Fiske program seems to hold. That is, _convergent validity_ seems to be logically necessary, when we modify the statement of this criterion. However, such modifications reduce us to circular reasoning.

(2) That the other criteria aimed at guananteeing _discriminant validity_ (II - IV) do not seem to be based on _a priori_ grounds. There does not seem to be anything in the very nature of testing which requires that tests be "discriminantly valid." This conclusion does _not_ imply that there are not any sound grounds for asking that tests be discriminantly valid. There may well be sound utilitarian grounds, but these are contingent, not necessary, and must be handled accordingly.

Part I Monte Carlo Analysis

1. Stanley's F statistic for determining trait validity by multitrait-multimethod matrices was not robust and was not invariant for non-null contributions of method and method-trait bias.

2. Conditions could be prescribed for using Stanley's F statistic, under non-null conditions of method(b_i) and method-trait(W_{jk}) bias. These conditions are presented in Table 11 and provide the best fit of the theoretical and empirical distributions under non-null conditions of these two sources of bias.

TABLE 11

Best Weightings of Method(b_j)
and Method-trait(W_{jk}) Bias.

M	T	b_j	W_{jk}	Remarks
2	2	8/100	92/100	For P=5, 100 neg. F's After P=10 and of ρ
2	3	25/100	75/100	Independent of P and of ρ
2	4	38/100	62/100	" " " " " "
3	3	not clear		Not clear, prob. around $\frac{4-6}{10}$
3	4	33/100	67/100	Independent of P, independent of ρ
4	4	not clear		Not clear, lowest X^2 at $\frac{1}{14}$ - $\frac{13}{14}$ but poor $\bar{\rho}_{emp}$
2	5	38/100	62/100	
3	5	44/100	56/100	
4	5	not clear		

Implications

The results of this study imply that there is a need for two-fold investigations of validity. Logical analyses which may clarify the concepts underlying testing and test validity are direly needed. This investigation scarcely scratched the surface of these problems yet it shook the foundations on which testing is based. [Empirical analyses using Monte Carlo techniques to evaluate the effectiveness of possible theoretical corrections of Stanley's F statistic could make this statistic and Campbell and Fiske's criteria for multitrait-multimethod validity more usable.]

The Monte Carlo analysis showed promise and prescribed conditions under which Multitrait-multimethod statistics could be useful in determining validity. This usefulness could be amplified with an expanded Monte Carlo analysis.

SUMMARY

This research investigated the appropriateness of using multitrait-multimethod intercorrelation matrices and Campbell and Fiske's criteria (1) as a validational process. This was a two part investigation utilizing a Monte Carlo analysis (Part I) and a Logical Analysis (Part II) and these are summarized separately.

Part I. Monte Carlo Analysis

The Monte Carlo analysis investigated the appropriateness of using the statistics developed for multitrait-multimethod intercorrelation matrices to validate data obtained from small sample sizes. These statistics were developed by Stanley (8) and Zyzanski (10) using three-way factorial designs where the three factors were persons, methods and traits.

The objectives of this part of the study were:

1. To generate for small sample sizes, empirical distributions of the F statistics (Stanley's and Zyzanski's) for testing trait validity in a multitrait-multimethod matrix.
2. To determine if these statistics remain invariant for various combinations of non-null contributions of the sources of method and error bias.
3. To compare Stanley's statistic with Zyzanski's and with the criteria of Campbell and Fiske.
4. If necessary, to present the prescribed conditions which permit the use of these statistics.

Objectives 1 and 2 were achieved by the following prodedures. The mathematical model for obtaining the Person-Method-Trait scores is given in equation 16.

$$(16) \quad X_{ijk} = P_i + b_j m_{ij} + w_{jk} e_{ijk}$$

In equation 16 the terms P_i, m_i; and e_{ijk} were random normal numbers generated on the computer, and represent null conditions as described in the Method chapter. The non-null conditions were represented by the terms b_j and W_{jk} which were treated as two weighting factors. P_i represented each persons variability. The other terms, b_j, m_{ij}, w_{jk} and e_{ijk} represented the four possible sources of method bias which are estimated by variance components attributable to : method (halo) effect (b_j), person-by-method interaction effect (m_{ij}), method-trait interaction effect (W_{jk}), and person-by-method-by-trait interaction effect (e_{ijk}).

The two weighting factors (b_j and w_{jk}) were related by restricting the average correlation over persons ($^{-}$ $x_{ijk}, x_{ij'k'}$) to three categories, low ($\bar{r} = .3$), medium ($\bar{r} = .7$), and high ($\bar{r} = .9$). Theoretically the weighting factors and the correlation are related by equation 17.

$$(17) \quad {}^{-}(x_{ijk}x_{ij'k'}) = \bar{1}\ \frac{1}{1 + b^2_j + w^2_{jk} \quad 1 + b^2_{j'} + w^2_{j'k'}}$$

The weighting factors were restricted to specific degrees of inequality and to specific proportions of total variance which they contributed and were determined for the three values of by means of equation 17.

Once the Person-Method-Trait, PMT, scores were obtained these were correlated over persons to give an MT by MT (M is the number of Methods and T the number of Traits) intercorrelation matrix. Both Stanely's and Zyzanski's F statistics for testing person-trait-interaction were calculated for this matrix using both adjusted (Zyzanski's) and unadjusted (Stanley's) correlation coefficients. The entire procedure for obtain this matrix and statistic were repeated 1000 times. This gave an empirical distribution with 1000 points for each statistic.

Stanley's F_{pt} Statistic

Approximately 150 such empirical distributions were generated. Each empirical distribution was compared with its theoretical F distribution with the chi-squared goodness of fit test. The results of these comparisons are given in tables 1 through 9.

Results

This research investigated the appropriateness of using multitrait-multimethod intercorrelation matrices and Campbell and Fiske's criteria (1) as a validational process. This was a two part investigation, statistical and logical, and these were treated separately and the results are reported separately.

The Monte Carlo analyses investigated the multitrait-multimethod intercorrelation matrices to validate data obtained from small sample sizes. These statistics were developed by Stanley (8) and Zyzanski (10) using three-way factorial designs where the three factors were persons, methods, and traits.

Inspection of Tables 1 through 9 reveals that Stanley's F statistic is not robust and is not invariant to non-null contributions of method and method-trait bias. Graphs 1 and 2 present data which show that is is possible to minimize the distorting effects of these non-null contributions of method and method-trait bias. Tables 10 and 11 summarize those conditions which prescribe the usefulness of Stanley's F statistic for small sample sizes.

Zyzanski's F statistic which can be considered a correction of Stanley's could not be generated satisfactorily without obtaining more than 10 per cent negative F values. It was concluded that Zyzanski's adjusted F statistic should not be used with small sample sizes.

It is recommended that other Monte Carlo analyses be made in order to expand the usefulness of Stanley's F statistic in the validation of data obtained from small sample sizes.

Part II. Logical Analysis

I. Purpose of our Investigation*

Personality-trait tests are widely used and are being produced in abundance. The question then arises, "Which tests are good or valid?" There ought to be a way to answer this query. Campbell and Fiske, in their article entitled, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," offered four criteria which a valid test must meet. The purpose of this study was to examine critically these four criteria to determine whether the criteria are sound. Our task was to determine, at least in part, what grounds Campbell and Fiske had for saying that their four proposed criteria must be met by any good test.

The nature of our inquiry must not be misunderstood. We are not developing any particular testing method, nor are we highhandedly encroaching on the domain of testing. Our study, so to speak, does not "advance" the field and methods of testing. Rather our investigation goes "backward," returns to the concepts which underlie the field of testing, and attempts to analyze these concepts and their logical inter-relationships. Our work is an essay in the foundations of testing and proceeds a priori, not empirically. We are dealing with the concepts on which testing rests, not the facts which testing uncovers. Consequently, whereas Campbell and Fiske work on criteria to be used in judging the worth of a test, we are concerned with considerations which will enable us to judge the value of the criteria themselves.

---

*(Note: This precis assumes the reader has read Campbell and Fiske's article).

II. Method of Inquiry

Our method must also be carefully distinguished. We did not proceed statistically, for example. Rather we employed the method of logical analysis so frequently used by contemporary English-speaking philosophers. Our method, then, is philosophical, not empirical. And this method must be distinguished from certain contemporary approaches such as Existentialism and Phenomenology. Nor is this procedure comparable to some philosophy of education approaches which are historical in character. The techniques we employed were those of linguistic analysis, conceptual analysis and symbolic logic. We proceeded from the concepts of test, validity, and reliability (both technical and non-technical concepts) to determine whether the Campbell-Fiske criteria followed *a priori*, and therefore with logical necessity, from these concepts.

Our procedure was to examine our common-sense notions, as well as the technical concepts, of test, validity and reliability, and, where possible, to transform our results into symbolic logic to make the conceptual properties and relations as clear as possible.

III. Conclusions

The conclusions of our inquiry are the following:

(1) That only criterion I of the Campbell-Fiske program seems to hold. That is, convergent validity seems to be logically necessary, when we modify the statement of this criterion. However, such modifications reduce us to circular reasoning.

(2) That the other criteria aimed at guananteeing discriminant validity (II - IV) do not seem to be based on *a priori* grounds. There does not seem to be anything in the very nature of testing which requires that tests be "discriminantly valid." This conclusion does not imply that there are not any sound grounds for asking that tests be discriminantly valid. There may well be sound utilitarian grounds, but these are contingent, not necessary, and must be handled accordingly.

(3) That Campbell and Fiske have put their finger on a crucial problem in testing and have raised stimulating and valuable questions. One thing they help point out is that there is not only much need for a sustained effert to determine whether given particular tests are valid, but also whether the criteria offered to do this job are themselves "valid."

Even the putatively "thorough" treatments of test-validity are decidedly lacking in thoroughness, logical rigor and conceptual clarity. The moral is evident: if the principles of testing are not clear, it is difficult to imagine that any satisfactory application of them can be made. In short, a *most* important factor of a highly influential aspect of contemporary education, *etc*., has been sadly neglected. There is need for much work.

REFERENCES
(and Notes)

Notes:

1. Donald W. Camptbell and Donald T. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin 56 (March, 1959) 81-105.
2. Ibid., p. 81
3. Ibid., p. 83
4. Ibid., p. 81
5. Ibid., p. 84
6. Ibid., p. 85
7. Ibid., p. 83
8. Ibid., p. 83
9. Ibid., p. 82
10. See Campbell and Fiske, op. cit., pp. 81 and 83 and quoted above on p. 52.
11. Ibid., p. 84
12. Ibid., p. 82
13. Ibid., p. 83
14. Ibid., p. 83
15. Ibid., p. 83

REFERENCES:

1. Campbell, D.J., and Fiske, D.W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological Bulletin, 1959, 56: 81-105.
2. Cochran, W.G., "The comparison of different scales of measurement for experimental results." Iowa State College Annals of Mathematical Statistics, 1943.

3. Guilford, J. P., Psychometric Methods (2nd Ed.) N.Y. McGraw-Hill, 1954.

4. Humphreys, L.G. "Note on the multitrait-multimethod matrix." Psychological Bulletin, 1960, 57: 86-88.

5. Lord, F.M. "A significance test for the hypothesis that two variables measure the same trait except for errors of measurement." Psychometrika, 1957, 22: 207-220.

6. Lord, F.M. "Problems in Mental Test Theory Arising From Errors of Measurement." J. Amer. Stat. Assoc. 1959: 472-479 (a).

7. Lord, F.M. "An appropriate T statistic for medical test theory," Psychometrika, 24: 283-302 (b).

8. Stanley, J.C., "Analysis of Unreplicated Three-way Classifications, with Applications to Rater Bias and Trait Independence." Psychometrika, 1961 26: 205-219.

9. Willingham, W.W. and Jones, M.B. "On the identification of halo through analysis of variance. "Educational Psychology of Measurement, 1958, 18: 403-407.

10. Zyzanski, S.J. 1962 "Analysis of variance applied to factors which do not have comparable scales." Unpublished Master's thesis, Iowa State University of Science and Technology.

APPENDIX I

Selection and Description of Random Normal Number Generator

The initial step was to obtain a working random normal number generator which gave satisfactory results with as much speed as possible. The library routine (#V0039, Computer Science Center, University of Virginia) which uses an eleven digit generator (cf. Handbook of Mathematical Functions, National Bureau of Standards, 1964, p. 953) gave satisfactory results but was somewhat slow at 30.4 millisecs/random normal number. This was said to have been checked out for second-order correlations. However, an article (Communications of the American Computer Machine, vol. 3, 1965) stated that this particular sequence contains a third order correlation and third order are necessary to this study. An eight digit random number generator was subsequently chosen. Mathematically,

$$x_{i+1} = (6065_8 \quad x_i) \text{ moo } 2^{25}$$ (American Computer Machine, vol.8, 1965).

This when used in connection with ACM Algorithm # (cf. Alderman) gave a generation rate of 6 millisecs/random normal number. The distributions produced by this routine were plotted and checked against the theoretical distribution and the following resulted: (cf. plots)

initializing integer	no. pts.	no. int.	chi sq.
11111111	2000	19	28.4
33333333	2000	19	36.7
55555555	2000	19	13.1
77777777	2000	19	21.3
99999999	2000	19	17.2
55555555555*	2000	28	22.0

* Noran (V0039) used from library

To minimize any interaction within the test scores,

$$X_{ijk} = P_i + b_j m_{ij} + w_{jk} e_{ijk} ,$$

P, m, & e were obtained from separate random sequences initialized at 55555555, 77777777, and 99999999, respectively.

APPENDIX II

Description of Method for Determining Factors b_j(Method) and W_{jk}(Method-Trait).

The weighting factors b_j and w_{jk} were obtained from equation (17).

$$(17)\ \rho\,(ijk,ij'k') = \sum_{I} \frac{1}{\sqrt{1 + b^2_j + W^2_{jk}}\ \sqrt{1 + b^2_{j'} + W^2_{j'k'}}}$$

At first, the simplification b = W = Z was made and equality was assumed throughout the matrix X_{ijk}. This gave

$$\bar{\rho} = \sum_{I} \frac{1}{1 + 2Z^2} \quad \text{and given } \bar{\rho}\text{, b and W}$$

could be determined. To produce inner variations within i, j, and k and between b and W linear scaling was used, e.g.,

$$\text{for } \frac{b}{W} = \frac{1}{2},\ \ b = Z, \text{ and } W = 2Z$$

To vary b within j (methods) it would be weighted so that the equality $(\sum_{j}^{m} b_j)/m = \bar{b} = Z$ was maintained. These methods were checked and gave average $\bar{\rho}_{\text{Emp.}}$ close to $\bar{\rho}_{\text{Theor.}}$ except for the lower range $\bar{\rho}_{\text{Theor.}} = 0.3$ where $\bar{\rho}_{\text{Emp.}}$ 0.4 approximately. This may be the fault of the random normal number generator in which approximations were used in the interests of conserving computer time.

In generating the empirical distributions, an exact method was used for standard deviations ranging from -3 to +3. For standard deviations which vary from $3 \leq / x / \leq 6$, a linear approximation was used. No numbers were generated with standard deviations larger than 6.

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

ERIC ACCESSION NO.

CLEARINGHOUSE ACCESSION NUMBER

RESUME DATE: 09 01 67
P.A.: -
T.A.: -

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☒
ERIC REPRODUCTION RELEASE? YES ☒ NO ☐

TITLE
The Use of the Computer to Generate Statistical Tables for the Study of Personality Traits: (A Monte Carlo and Logical Analysis of Multitrait-Multimethod Statistics and Criteria for Validation)

PERSONAL AUTHOR(S)
Jacobson, Milton D.

INSTITUTION (SOURCE)
University of Virginia, Charlottesville, Va., School of Education

SOURCE CODE

REPORT/SERIES NO.

OTHER SOURCE
Office of Education, U.S. Dept. of Health, Education, Welfare

SOURCE CODE: 5-8410

OTHER REPORT NO.

OTHER SOURCE

SOURCE CODE

OTHER REPORT NO.

PUB'L. DATE: 1 Sept., 67
CONTRACT GRANT NUMBER: 5-8410

PAGINATION, ETC.
100 p.

RETRIEVAL TERMS
The Use of the Computer to Generate Statistical Tables for the Study of Personality Traits: (A Monte Carlo and Logical Analysis of Multitrait-Multimethod Statistics and Criteria for Validation)

IDENTIFIERS

ABSTRACT

This research investigated the appropriateness of using multitrait-multimethod intercorrelation matrix F statistics and criteria as a validational process. Monte Carlo Analyses of the robustness of these statistics were made. Logical analyses of the criteria were also made. It was concluded that the F statistics were not robust, but tables are presented which permit their use under prescribed conditions. It was also concluded that only criterion 1 (convergent validity) seems to be logically necessary while criteria 2-4 (discriminant validity) are severely weakened as they must be justified on utilitarian and contingent but not necessary grounds.